



Developing an Academic Collocation List for Arts and Humanities

Author: James O'Flynn

University of Warwick

British Council ELT Master's Dissertation Awards 2019

Commendation

Developing an Academic Collocation List for Arts and Humanities

James O'Flynn

1766500

Dissertation submitted in partial fulfillment of the requirements for the degree of
MA in ELT
(specialism in ESP/ EAP)

**Centre for Applied Linguistics
University of Warwick
September 2018**

Acknowledgements

First and foremost, I would like to thank my dissertation supervisor, Sue Wharton, for helping me make and justify sound practical and theoretical decisions throughout this challenging research project.

I would also like to thank all faculty members in the Centre for Applied Linguistics for their guidance and support during the dissertation research process, particularly Andy Davidson, Gerard Sharpling, Nigel Prentice, Sal Consoli and Tilly Harrison for their participation.

Further thanks go to Mark and Alison O'Flynn, without their generous hospitality MA study would not have been possible, and to Katie Webb, who (without complaining too much) listened to me jabber on about the multifaceted phenomenon of collocation – something which is of very little interest to her.

Finally, I would like to acknowledge the Sketch Engine team at Lexical Computing Ltd. for their swift responses to any and all of my technical questions, with particular thanks to Michal Cukr for his patience.

Abstract

Corpus-derived lists of academic vocabulary are widely used in the teaching-learning of English for Academic Purposes (EAP). The most noteworthy of these lists are the Academic Word List (Coxhead, 2000), the Academic Collocation List (Ackerman and Chen, 2013) and the Academic Formulas List (Simpson-Vlach and Ellis, 2010). Each of these lists is based on the assumption that there exists a 'core' academic vocabulary which is equally useful across all academic disciplines. Yet, there is mounting evidence to suggest that academic vocabulary occurs and behaves in dissimilar ways in different disciplinary environments (Hyland and Tse, 2007; Hyland, 2008; Durrant, 2009). In light of this, there have been numerous attempts to compile discipline-specific academic word lists for EAP (e.g. Wang, Liang and Ge, 2008; Martinez, Beck and Panza, 2009, Vongpumivitch, Huang and Chang, 2009; Li and Qian, 2010), yet there have been no attempts to compile comparable lists of multiword units. This dissertation, therefore, aims to demonstrate the need for an empirically-derived discipline-specific list of academic collocations for EAP, moreover ESAP, and then presents and evaluates such a list – the Academic Collocation List for Arts and Humanities (ACLAH). The results suggest that the ACLAH represents progress towards a more comprehensive account of academic collocation which would better serve EAP students in Arts and Humanities than lists of 'generic' academic vocabulary. The study concludes with a series of implications for both future research and teaching-learning with the ACLAH.

Developing an Academic Collocation List for Arts and Humanities

Table of contents

1. Introduction	1
2. Literature Review	3
2.1. The phraseological tendency of language	3
2.2. The importance of collocation in SLA and EAP	4
2.2. Approaches to collocation	5
2.2.1. The neo-Firthian approach	5
2.2.2. The phraseological approach	7
2.2.4. Approach to collocation in the present study	8
2.3. Previous attempts to compile vocabulary lists for EAP teaching-learning	9
2.3.1. Lists of single-word academic vocabulary	9
2.3.2. A list of academic formulas	10
2.3.3. Lists of academic collocations	11
2.3.4. Lists of discipline specific academic vocabulary	13
2.3.5. Approach to compiling a list of academic vocabulary in the present study	14
3. Methodology	15
3.1. Corpus compilation	15
3.1.1. Compiling a preliminary corpus based on a university disciplinary map	16
3.1.2. Performing a keyword analysis	18
3.1.3. Restructuring the preliminary corpus to produce the final corpus	21
3.2. Computational analysis of the AHC to produce a preliminary list of academic collocations	22
3.2.1. Search attribute: collocations	23
3.2.2. Span: +/-5	24
3.2.3. Frequency: 28	25
3.2.4. Keyness: add-n threshold of 0.1	25
3.2.5. Results from the computational analysis	27

3.3. Manual refinement of syntactic combinations	28
3.4. Manual removal of noise	30
3.5. Computational retrieval of dispersion values	33
3.6. Manual refinement based on degree of fixedness	34
3.7. Computational retrieval of statistical significance values	35
3.8. Manual refinement by expert review.....	37
3.8.3. Preparing the expert review	37
3.8.2. Results from the expert review	39
3.9. Presenting the ACLAH	44
4. Results and discussion	46
4.1. Composition of the ACLAH	46
4.2. Validation	50
5. Conclusions and implications.....	55
References	59
Appendices.....	66
Appendix 1. Overview of the study and the subset of 112 collocations	66
Appendix 2. The Academic Collocation List for Arts and Humanities	73

List of Abbreviations

EAP – English for Academic Purposes

ESAP – English for Specific Academic Purposes

AWL – Academic Word List (Coxhead, 2000)

AFL – Academic Formulas List (Simpson-Vlach and Ellis, 2010)

ACL – Academic Collocation List (Ackerman and Chen, 2013)

L2 – Second Language

ACLAH – Academic Collocation List for Arts and Humanities

AH – Arts and Humanities

AHC – Arts and Humanities Corpus

ENGCOMP – English and Comparative Literary Studies

FILMTV – Film and Television Studies

MODLANG – Modern Languages and Cultures

THEATRE – Theatre and Performance Studies and Media and Cultural Policy Studies

ARTHIST – History of Art

HIST – History (including American Comparative Studies)

POS – Part of speech

LDOCE – Longman Dictionary of Contemporary English

ICC – Intraclass Correlation Coefficient

PCC – Pearson Correlation Coefficient

BAWE – British Academic Written English corpus

POS Suffixes

-a (e.g. *specifically-a*) adverb

-j (e.g. *previous-j*) adjective

-v (e.g. *shoot-v*) verb

-n (e.g. *research-n*) noun

1. Introduction

Corpus-derived lists of academic vocabulary are widely used in the teaching-learning of English for Academic Purposes (EAP). This is because academic vocabulary is neither sufficiently frequent in language to be learned implicitly nor likely to be taught explicitly as part of subject courses (Nation, 2001:189-191). Empirically-derived lists of academic vocabulary, then, are extremely useful and can form the basis of a lexical EAP syllabus (Laufer, 1991; Laufer and Sim, 1985, Nation and Hwang, 1995; Nation and Waring, 1997, Ward, 1999). The most commonly-used of such listings is the Academic Word List (Coxhead, 2000), which, upon publication, was commended for being the most extensive investigation of academic vocabulary to date (Hyland and Tse, 2007). Gradually, though, corpus-based research interest in academic vocabulary has shifted away from single words and towards co-occurring words (e.g. Biber, Conrad and Cortes, 2004; Simpson-Vlach and Maynard, 2008), resulting in a number of lists of multiword units for EAP teaching-learning purposes. The most noteworthy of these lists are Durrant's listing of academic collocations (2009), the Academic Collocation List (ACL) (Ackerman and Chen, 2013) and the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010).

The focus of Durrant's listing and the ACL are two-word collocations, while the focus of the AFL is 3-, 4- and 5-word formulas. Beyond just the number of co-occurring words comprising a multiword unit, each list is different to the other for various reasons. For example, Durrant's (2009) strictly computer-based approach resulted in a final listing of 1000 predominantly *grammatical collocations*, while Ackerman and Chen's (2013) mixed-method approach, fusing computational analysis with human intervention, resulted in a final ACL of 2,468 *open and restricted lexical collocations*. Yet, all three lists share one very significant similarity – they are all based on the assumption that there exists a 'core' academic vocabulary which is equally useful across all academic disciplines.

In two prominent corpus-based studies, Hyland and Tse (2007) and Hyland (2008) questioned the assumption of a 'core' academic vocabulary and found that both single words and multiword units occur and behave in dissimilar ways in different disciplinary environments. Both studies concluded that the best way to prepare students for their studies is to provide them with an understanding of the language features of their particular courses. In light of this, there have been numerous attempts to compile discipline-specific academic word lists (e.g. Wang, Liang and Ge, 2008; Martinez, Beck and Panza, 2009, Vongpumivitch, Huang and Chang, 2009; Li and Qian, 2010), yet there has hitherto been no attempt to compile a list of discipline-specific multiword units for EAP. This dissertation, therefore, aims to demonstrate the need for an empirically-derived discipline-specific list of academic collocations for EAP, moreover

English for Specific Academic Purposes (ESAP), and then presents and evaluates such a list – the Academic Collocation List for Arts and Humanities (ACLAH).

2. Literature Review

The aim of this chapter is to provide a background of the present study and highlight the need for a list of discipline-specific academic collocations. In the following subsections, a number of key notions and will be explored: the phraseological tendency of language (2.1), collocation in SLA and EAP (2.2), approaches to collocation (2.3), and vocabulary lists for EAP teaching-learning (2.4).

2.1. The phraseological tendency of language

Multiword units are currently viewed as a necessary component of lexical competence in SLA (Laufer and Waldman, 2011:648). This is because, as argued by Pawley and Syder (1983:215), 'by far the largest part of the English speaker's lexicon consists of complex lexical items including several hundred thousand lexicalised sentence stems'. In other words, 'natural language makes considerable use of recurrent formulaic patterns of words' (Simpson-Vlach and Ellis, 2010:373). It is, therefore, extremely rare for a language user to have complete freedom of choice of a single word. In fact, as Sinclair (1991:110) posits in his *Idiom Principle*, 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices'. It seems, then, that the most basic units of language are not words but recurrent constructions which, as single phraseological choices, reduce cognitive effort, save processing time, and render language available for immediate use, improving both the quality and fluency of spoken and written language (Pawley and Syder, 1983; Shin and Nation, 2008; Ellis, 2009).

In EAP, multiword units are a particularly important aspect of lexical competence because corpus based-studies confirm that they are not only salient but also functionally significant in academic discourse (Simpson-Vlach and Ellis, 2010:487). Biber, Conrad and Cortes (2004), who examined 4-word lexical bundles in a corpus of academic speaking and writing, found that these multiword units are associated with particular semantic, pragmatic and discourse functions and thus are a fundamentally important part of academic writers' and speakers' communicative repertoire. What is more, in another corpus-based study, Ellis, Simpson-Vlach and Maynard (2008) examined the processability of 3- 4- 5-word formulas in academic discourse and concluded that L2 learners clearly need support in learning these multiword units and therefore EAP instruction should seek to identify and prioritise which formulas to teach. These two studies, which had wide-ranging implications for the teaching-learning of lexis in EAP, shifted the focus from single words to co-occurring words.

2.2. The importance of collocation in SLA and EAP

Collocation, in its broadest sense, refers to the habitual co-occurrence of a word with another word or words with a greater frequency than would be expected by chance - a phraseological phenomenon which is prevalent in natural language use. This prevalence makes high-frequency collocations a 'part of the lexicon which learners need to acquire' (Durrant, 2009:158). Lewis (1993) first stressed the importance of teaching collocations in his influential Lexical Approach, and later presented an edited volume with many innovative ways to teach collocations (see Lewis, 2000). Yet, for L2 learners, achieving a high level of collocational competence is not easy. For example, as Shin and Nation (2008:340) point out, Korean students drawing on their first language are likely to collocate *artificial* with *teeth* for *false teeth* and *thick* with *tea* for *strong tea*. Nesselhauf (2005) suggests that 50% of collocation errors (e.g. *thick tea*) are due to L1 interference. This can be explained by Hoey's (2005:8) theory of Lexical Priming:

'As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context.'

That is to say, when a native speaker perceives or produces the word *tea*, he or she is psychologically primed to perceive or produce one or more of the words *tea* is cognitively associated with, for example *strong*¹, whereas a non-native speaker might be more likely to perceive or produce *thick* or *powerful*, depending on their first language.

It is breaking these primings and establishing new ones which is challenging for L2 learners. As Laufer (2011: 30) states, 'the use of collocation is problematic for L2 learners, regardless of years of instruction they received in L2, their native language, or type of task they are asked to perform'. In fact, Bahns and Eldaw (1993) suggest that collocations are particularly problematic for advanced learners because collocational competence does not develop in parallel with general vocabulary competence. Many L2 speakers, therefore, over-rely on a small number of collocations (Cobb, 2003), which can be problematic in academic environments where collocation is ubiquitous and highly discipline-specific (Ward, 2007:21).

¹It is, of course, possible to establish new primings or override existing ones. For example, in literary work, the intentional breaking of traditional primings (e.g. *powerful tea*) enables writers to be creative (see Hoey, 2007).

As Gledhill (2000:1) asserts, 'it is impossible for a writer to be fluent without a thorough knowledge of the phraseology of the particular field he or she is writing in'. It is, then, especially important that collocations are taught explicitly as part of EAP courses on which one of the key aims is to develop more fluent and accurate levels of academic language use (Storch and Tapper, 2009:208).

2.2. Approaches to collocation

There is no simple and precise definition of collocation. As Bahns (1993:57) states, 'collocation is a term which is used and understood in many different ways'. This heterogeneity of definition is the result of multiple approaches to the phenomenon of collocation. Due to the constraints of this paper, only the two most relevant to the present study will be reviewed, they are: the neo-Firthian approach; and the phraseological approach. Although these two approaches define and operationalise the term collocation somewhat differently, they are not to be considered contradictory.

2.2.1. The neo-Firthian approach

Collocation is an old idea brought to its modern form by Firth (1957), who famously coined the adage 'you shall know a word by the company it keeps' (ibid:11). This conception of collocation as 'an abstraction at the syntagmatic level' (ibid:196) became the impetus for a new school of corpus linguistics, the neo-Firthian approach, to which the notion of collocation is central. For neo-Firthians, collocation, at its simplest, is 'the occurrence of two or more words within a short span of each other in a text' (Sinclair, 1991:170). There are, though, two techniques for identifying collocations within this school of corpus linguistics. McEnery & Hardie (2012) refer to these two techniques as *collocation-via-concordance* and *collocation-via-significance*.

Collocation-via-concordance was frequently adopted in the earlier stages of corpus linguistics and is largely associated with Sinclair's (1966:415) definition of collocation:

'We may use the term node to refer to an item whose collocations we are studying, and we may define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call collocates.'

By this definition, any co-occurrence within a set span is considered a collocation. It is the role of the computer to simply supply concordance lines (hence *collocation-via-concordance*) and

the role of the linguist to examine them individually for recurring items and patterns which they deem significant. Here, then, significance is judged intuitively rather than statistically. This technique has been utilised by neo-Firthians to expand the notion of collocation to more abstract concepts such as colligation (Hoey, 2005; Sinclair, 1996, 1998, 2004), semantic preference (Sinclair, 1991; Stubbs, 1995) and semantic prosody (Louw, 1993; Sinclair, 1991; Stubbs, 1995, 1996, 2001). Because these concepts have become central to neo-Firthian corpus linguistics, there is a tendency among neo-Firthians to favour the collocation-via-concordance technique (McEnery & Hardie, 2012:130).

In contrast, *Collocation-via-significance*, which is of most relevance to the present study, does not identify collocations exclusively on the basis of co-occurrence within a certain span, but rather whether or not the co-occurrence within a certain span is statistically significant. This technique for identifying collocations is clearly central to later definitions of collocation:

‘The test of whether two words are significant collocates... requires 4 pieces of data; the length of the text in which the words appear, the number of times they both appear in the text, and the number of times they occur together’ (Sinclair, Jones & Daly’s, 2004:28).

The measure of significance in this definition is clearly concerned with mathematical evidence that items co-occur ‘with greater than random probability’ (Hoey, 1991:7). This mathematical evidence can be acquired using statistical significance tests such as t-score, mutual information (MI) and/or logDice. The choice of statistical measure, though, requires careful consideration because, as Hunston demonstrates (2002:61-71), different measures prioritise different aspects of collocation which has a major effect on determining what is and what is not a collocate (Table 1). For example, a t-score will vary depending on the size of a corpus while an MI score will favour low frequency collocations². Some researchers, therefore, argue that frequency-based statistical measures *alone* are not a reliable criterion for identifying significant collocations (e.g. Kjellmer, 1987), which is why, even with collocation-via-significance, ‘the researcher is still regarded as the final arbiter of determining whether or not a specific candidate is indeed a collocate’ (McEnery & Hardie, 2012:126).

Table 1. The top five collocates of ‘research’ within a span of +/-4 in the BAWE

T-score	MI	logDice
.	disseminating	qualitative

² Statistical significance measures will be discussed in more detail in section 3.7, where they are of most importance to the present study.

,	rigour	journal
of	trustworthy	methods
the	feed-forward	into
and	nexo	further

2.2.2. The phraseological approach

The phraseological approach is largely concerned with classifying and demarcating collocations according to their varying degrees of fixedness. Within this approach, collocations fall on a continuum from the most restricted to the most free. The markers along the continuum have been given various labels by various phraseologists (Table 2), but generally three classifications are distinguishable: idioms at the most fixed extreme, open collocations at the most free extreme, and restricted collocations in the centre. According to Cowie (1998:5), restricted collocations are the most interesting, yet the most difficult to demarcate because they are somewhere between open collocations and idioms on the continuum. In order to distinguish what is free and what is fixed, three criteria are typically applied: semantic transparency; specialised use of one component; and commutability.

Table 2. *The Continuum Model*

Phraseologist	Labels (from most fixed to most free)			
Cowie (1981)	Pure idiom	Figurative idiom	Restricted collocation	Open collocation
Nattinger and Decarrico (1992)	Idiom		collocation	Free combination
Howarth (1998)	Pure idiom	Figurative idiom	Restricted collocation	Free combination

The first criterion, semantic transparency, which distinguishes idioms from open and restricted collocations, refers to ‘whether meaning attaches to the whole or to the parts of a unit’ (Howarth, 1996:38). In the word combination *kick the bucket*, for example, the meaning (*to die*) is not the sum of its constituent parts, therefore it displays a high degree of fixedness and is classified as an idiom. To distinguish a restricted collocation from an open collocation, then, two further criteria are applied: specialised use of one component and commutability. That is to say, in a restricted collocation, at least one component must have a non-literal (specialised) meaning and at least one a literal one, and commutability must be arbitrarily restricted (though some commutability is possible) (Nesselhauf, 2005:25). In the collocation *adopt (a) policy*, for example, the word *adopt* is used in a figurative sense (i.e. specialised sense), and the number of alternative direct objects of *adopt* when used in the sense of *start to use* is arbitrarily

restricted. Therefore, *adopt a policy* fulfils the two aforementioned criteria and is classified as a restricted collocation (Table 3). In sum, restricted collocations must be semantically transparent, have a specialised component and be arbitrarily restricted.

Table 3. Examples of word combinations classified using the phraseology approach

Idioms	Restricted collocations	Open collocations
<i>Kick (the) bucket</i>	<i>Adopt (a) policy</i>	<i>Write (an) essay</i>
<i>Face (the) music</i>	<i>Conduct research</i>	<i>Cashier (an) officer</i>
<i>Spill (the) beans</i>	<i>Commit (a) crime</i>	<i>Broken window</i>

While idioms are not manipulated at all, restricted collocations are manipulated as phraseological units and open collocations are *generally* manipulated based on grammatical rules. However, both Howarth (1996:41-42) and Ackerman and Chen (2013:236) acknowledge that in open collocations no single word can genuinely co-occur with any other without some level of arbitrary restriction. For example, in the collocation *cashier (an) officer*, *cashier* is used in its literal sense, making it an open collocation, but *cashier* as a transitive verb is arbitrarily restricted to the object noun *officer* or another rank in the armed forces (Howarth, 1996:42). Therefore, open collocations, like restricted collocations, are also subject to varying degrees of arbitrary restriction. It is this arbitrary restriction which makes both open and restricted collocations challenging for learners to master (Ackerman and Chen, 2012:239), and for this reason, these two classifications of collocation are of most interest to the present study.

2.2.4. Approach to collocation in the present study

Based on the aforementioned literature, the present study regards collocations as statistically significant word pairs which are subject to varying degrees of arbitrary restriction. This definition borrows from two approaches - the neo-Firthian approach, moreover collocation-via-significance, and the phraseological approach. The former approach offers the means by which to quantitatively identify a large number of mathematically significant collocations using computer software, while the latter approach offers a model for qualitatively refining a long list of collocations based on their degree of fixedness (among other things) in order to ensure that the final list is pedagogically relevant.

2.3. Previous attempts to compile vocabulary lists for EAP teaching-learning

It has long been known that ‘the language necessary for proficiency in academic contexts is quite different from that required for basic interpersonal communicative skills’ (Simpson-Vlach and Ellis, 2010:487). For this reason, there have been numerous attempts to compile lists of single-word and multiword vocabulary for EAP teaching-learning purposes. Those attempts which are most relevant to the present study will be reviewed in the following subsections: lists of single-word academic vocabulary (2.3.1); a list of academic formulas (2.3.2); lists of academic collocations (2.3.3); and lists of discipline specific academic vocabulary (2.3.4).

2.3.1. Lists of single-word academic vocabulary

Research interest in lists of academic collocations was to some degree borne out of the many attempts to compile lists of single-word academic vocabulary (e.g. Champion and Elley, 1971; Praninskas, 1972; Lynn, 1973; Ghadessey, 1979; Xue and Nation, 1984; Coxhead, 2000; Pacquot, 2010; Gardner and Davies, 2014; Browne, Culligan and Phillips, 2014). The most commonly-used of these lists is arguably Coxhead’s (2000) Academic Word List (AWL), a list of 570 word families that cover, on average, 10% of any academic text (Fig. 1). Upon publication, the AWL ‘filled a substantial gap in language education by providing a corpus-based list of lexical items targeted specifically for academic purposes’ (Simpson-Vlach and Ellis, 2010:489).

Fig. 1. Word families in the Academic Word List

analyse	approach	assessable	benefit	concept
analysed	approachable	assessed	beneficial	conception
analyser	approached	assesses	beneficiary	concepts
analysers	approaches	assessing	beneficiaries	conceptual
analyses	approaching	assessment	benefited	conceptualisation
analysing	unapproachable	assessments	benefiting	conceptualise
analysis		reassess	benefits	conceptualised
analyst		reassessed		conceptualises
analysts		reassessing		conceptualising
analytic		reassessment		conceptually
analytical		unassessed		
analytically				

The AWL, then, was generally lauded by EAP teachers and researchers alike, but it is not entirely problem free. It has been criticised for breaking into single words units which may be better learnt as wholes and thus ignoring the collocational behaviour of words. Hyland and Tse (2007:247), for example, point out that *strategy* is often found in *marketing strategy* in business texts, *learning strategy* in applied linguistics texts and *coping strategy* in sociology texts (ibid:246). They view this divergence in collocational behaviour across disciplines as a key factor undermining the notion of a ‘core’ academic vocabulary and in turn the AWL (ibid:251). Furthermore, many high-frequency words which serve technical functions in academic collocations are often ignored in EAP instruction because the AWL omits the most frequent 2000 words of English, as represented by West’s (1953) General Service List (GSL). Durrant (2009:164), for example, points out that technical uses of high-frequency GSL items such as *address* in *address (an) issue*, might be overlooked in EAP because they are not on the AWL. Perhaps for reasons such as these, even Coxhead (2008) herself acknowledged that it was necessary to extend existing word lists to take account of multiword units.

2.3.2. A list of academic formulas

The Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010:487) addressed the gap in existing vocabulary lists by providing a pedagogically useful list of multiword units in the form of formulaic 3- 4- and 5-word sequences. The AFL, derived from a 4.2 million word corpus of general academic discourse, comprises 207 spoken formulas, 207 written formulas and 207 ‘core’ formulas. The formulas are prioritised by pedagogical relevance, as determined by a panel of EAP practitioners. This highlights an important difference between AFL formulas, which are prioritised perceptually by EAP practitioners, and other multiword units, such as lexical bundles, n-grams and clusters *inter alia*, which are *generally* prioritised based on frequency alone. For example, units like *to do with the* might rank highly on a strictly frequency-based list of lexical bundles, whereas intuition-based weeding and ranking ensures the AFL prioritises psycholinguistically salient formulas, e.g. *on the other hand* (Fig. 2). This is particularly important in EAP because ‘lists of highly-frequent expressions are of minimal use to instructors who must make decisions about what content to draw students’ attention to for maximum benefit’ (Simpson-Vlach and Ellis, 2010:490). For this reason, the present study will incorporate an intuition-based review by a panel of experts in order to remove and rank entries based on their pedagogical relevance (see 3.8).

Fig. 2. The most highly-ranked formulas in the ‘core’ AFL

in terms of	as a result of	whether or not
-------------	----------------	----------------

at the same time	this is a	the same time
from the point of view	on the basis of	with respect to
in order to	a number of	point of view of
as well as	there is no	as a function of
part of the	point of view	at the same
the fact that	the number of	the point of view
in other words	the extent to which	in such a way
the point of view of	as a result	the use of
there is a	in the case of	in other words the

The AFL represents a significant development in EAP vocabulary research and resources, shifting the focus from single-word vocabulary to multiword units, yet it is important to note that formulas, unlike collocations, are highly-fixed sequences with very little variation of individual components. This means the AFL leaves out much that might be of collocational interest in EAP because collocation ‘often involves relationships between words which may be separated by other, non-fixed, or semi-fixed words, and which may differ in their position to one another’ (Durrant, 2009:158).

2.3.3. Lists of academic collocations

To address the need for a list of positionally variable multiword units, there have been two attempts to compile lists of academic collocations, each taking a very different approach. The first attempt was by Durrant (2009:159), whose approach was purely frequency-based, i.e. based on statistics alone without any level of human intervention. He defined collocation in neo-Firthian terms, moreover collocation-via-significance, as word pairs that co-occur within a four-word span with a minimum normed frequency of 1 p/m words and a minimum MI score of 4 in each subcorpus of his 25-million-word general academic corpus. Furthermore, to be considered ‘academic’, collocations had to appear significantly more frequently in his general academic corpus than in a non-academic reference subcorpus of the BNC, calculated using Scott’s (1999) log-likelihood-based ‘keyword’ technique³. This technique for identifying ‘academic’ collocations will be used in the present study (see 3.2.4) because it does not have

³ The keyword technique is discussed in more detail in the methodology, section 3.2.4.

the inherent disadvantages of the other commonly used technique, which entails excluding items from the outdated GSL (West, 1953).

Durrant's (2009) final list includes 1000 academic collocations⁴, yet over 75% of the collocations on his list are 'grammatical collocations' (Benson, 1985), that is, the combination of one closed-class component (aka function word) and one open-class component (aka content word), e.g. pronoun + noun (*our study*) (Fig. 4). These grammatical collocations consist of relatively fixed and predictable patterns which are easier for learners to acquire (Ackerman and Chen, 2013:246), which is why they are not the typical focus of collocation studies. In fact, even Durrant (2009:165), who argues that grammatical collocations *are* legitimate EAP learning targets, concedes that such a high percentage of them is a caveat of his listing. For this reason, the present study will actively exclude grammatical collocations from the final list (see 3.3).

Fig. 4. Grammatical collocations from Durrant's listing

This study	Due to	Compared to	These results
Associated with	Consistent with	Was used	Respect to
This paper	Between and	Present study	To determine
Based on	Was performed	Number of	Note that
And respectively	Related to	As shown	Our study

The Academic Collocation List (2013), the second attempt to compile a list of academic collocations, is derived from a 25.6 million words corpus of general academic English and comprises 2,468 open and restricted lexical collocations (Fig. 5). Ackerman and Chen (2013), the compilers of the ACL, combined computational analysis with human intervention, arguing that Durrant's (2009) method, 'based on statistics alone, does not provide readily usable materials for EAP teaching' (ibid:236). Therefore, although they define collocation in neo-Firthian terms as words that co-occur within a three-word span with a minimum MI score of 3 and a minimum t-score of 2, they also use human intervention (like Simpson-Vlach and Ellis, 2010) to ensure that 'final entries are appropriate and relevant for EAP' (Ackerman and Chen, 2013:246). Human intervention allowed them to apply qualitative criteria typical of the phraseological approach. They were able to manually target exclusively 'lexical collocations' (Benson, 1985), that is, the combination of two open-class components, e.g. *verb + noun* combinations (*perform [an] experiment*), which are more difficult for learners to master than

⁴ Only the 100 most key collocations are available, as published in Durrant's (2009) paper – a full-list was never made available.

grammatical collocations (i.e. Durrant’s collocations) (Ackerman and Chen, 2013:246). Furthermore, they were able to manually target exclusively open and restricted collocations, which, as previously discussed, are subject to varying degrees of arbitrary restriction making them challenging for learners to acquire (Ackerman and Chen, 2012:239). The present study will adopt their techniques for targeting open and restricted lexical collocations (see 3.3 and 3.6).

Fig. 5. Lexical collocations from the ACL

verb/ noun	adjective/ noun	adverb/ adjective	adverb/ verb	noun/ noun
achieve goal	academic writing	ever increasing	closely rooted	background knowledge
achieve objective	brief overview	hardly surprising	generally accepted	class consciousness
cast doubt	causal link	adversely affect	well established	conflict resolution
make living	conflicting interests	mutually exclusive	differ significantly	data set
make prediction	conventional wisdom	radically different	expand rapidly	source material

The mixed-method approach to compiling the ACL, which included a full review of all entries by a panel of EAP experts, renders it more relevant and readily usable than Durrant’s (2009) listing, yet the ACL is not entirely problem free. In order to identify ‘academic’ collocations, Ackerman and Chen (2013:237) disallowed General Service List (West, 1953) items from occurring as node words in the ACL. This means that, although they allowed GSL items to occur pre- or post-node, an ACL collocation cannot be a combination of two GSL items, e.g. *control group*, and therefore certain collocations that could be of interest in EAP might be overlooked by the ACL. It is for this reason that the present study adopts Durrant’s (2009) approach to identifying ‘academic’ collocations through ‘keyness’ (see 3.2.4).

2.3.4. Lists of discipline specific academic vocabulary

The AWL (2000), the AFL (2010), Durrant’s listing (2009) and the ACL (2013) are all based on the assumption that there exists a *general* academic vocabulary common to all academic disciplines. There is, though, mounting evidence to cast doubt on this assumption. Hyland and Tse (2007) were the first to question the usefulness of generic academic vocabulary lists, finding that items in the AWL occur and behave differently across disciplines in terms of range, frequency, collocation and meaning. In another paper, Hyland (2008) found that multiword units also ‘occur and behave in dissimilar ways in different disciplinary environments’ (ibid:20). Interestingly, while Hyland and Tse’s (2007) paper set in motion a number of attempts to compile lists of discipline-specific single-word vocabulary (e.g. Wang, Liang and Ge, 2008; Martinez, Beck and Panza, 2009, Vongpumivitch, Huang and Chang, 2009; Li and Qian, 2010),

Hyland's (2008) paper failed to steer future research towards discipline-specific listings of multiword units.

The only list-compiler who has made a genuine attempt to address disciplinary variation in multiword vocabulary usage is Durrant (2009). Despite compiling a list of generic academic collocations, he examined the degree to which items on his listing were equally useful across disciplines. He found that in four of the disciplines represented in his general academic corpus (Life Sciences, Science and Engineering, Social-Administrative and Social-Psychological) the collocations in his list occurred between 30-35,000 times p/m words, while in the fifth discipline (Arts and Humanities) the occurrence rate was far lower at 17,677 occurrences p/m words. He concludes, therefore, that the vocabulary needs of students in Arts and Humanities (AH) should be treated separately from those of other EAP students (ibid:165). Yet, to the researcher's knowledge, there have hitherto been no attempts to compile discipline-specific lists of collocations for EAP. The present study, therefore, aims to fill this gap in vocabulary list research by producing an academic collocation list specifically for EAP in AH.

2.3.5. Approach to compiling a list of academic vocabulary in the present study

As previously stated, the present study approaches collocations as statistically significant word pairs which are subject to varying degrees of arbitrary restriction. In light of the aforementioned attempts to compile lists of academic vocabulary, this approach can now be further expanded upon. First, computational analysis will be used to generate a preliminary list of academic collocations, that is, collocations that are statistically more significant in an AH corpus than a non-academic reference corpus. Following this analysis, all entries will be manually filtered by the researcher to ensure that only the most pedagogically challenging classifications of collocation remain on the list, they are, open and restricted lexical collocations. Finally, the list will be vetted by a panel of experts to weed the list of pedagogically questionable entries and rank remaining entries by pedagogical relevance. The ACLAH will then be evaluated to address the following questions:

- (1) How is the ACLAH similar and/ or different to the ACL?
- (2) Is the ACLAH equally useful across Arts and Humanities fields?
- (3) Is the ACLAH a readily usable resource for EAP teaching-learning purposes?

3. Methodology

This chapter describes the development of the Academic Collocation List for Arts and Humanities (ACLAH). This involved many processes which are subsumed under three broad stages: corpus compilation (3.1); computational analyses (3.2, 3.5 and 3.7); and manual refinement (3.3, 3.4, 3.6, and 3.8).

3.1. Corpus compilation

In order to compile the corpus, Arts and Humanities must be first be defined, yet this is not an altogether straightforward task. Much has been written regarding the difficulty of defining academic disciplines. As Becher (1989:19) states that ‘the concept of academic discipline is not altogether straightforward, in that, as is true of many concepts, it allows for room for some uncertainties of application’. This is supported by Hyland (2012:23), who suggests that disciplines ‘have been seen in numerous ways: as institutional conveniences, networks of communication, political institutions, domains of values, modes of inquiry and ideological power-bases’. In addition to uncertainties of application, drawing disciplinary boundaries is made difficult by interdisciplinarity. Eldridge (2008:110), for instance, argues that ‘the academic environment and community is interdisciplinary and diffuse by nature’, perhaps because, as suggested by Hyland (2012:23), ‘research problems and investigations often ignore disciplinary boundaries’. Applied linguists, for example, are increasingly turning to mathematics and statistics to explain phenomena such as collocation. For these reasons, there are myriad ways by which disciplines can be defined.

One such way, and perhaps the simplest, is to use the disciplinary map of a university. Yet, Becher (1989:19) notes that there are theoretical issues with this approach such as how academic institutions elect to draw the map of knowledge. This study, for instance, requires a linguistically oriented approach to drawing the map of knowledge, but, as Durrant (2009:159) points out, ‘the university structure is obviously not based on linguistic decisions’. It is, though, possible to define disciplines linguistically, because, as Hyland (2012:11-32) suggests, we use language to identify and represent ourselves as legitimate members of a discourse community, and this community offers a way of bringing texts together into a common rhetorical space. Therefore, if it can be determined which discourse communities writers are orientating to and associating with, a linguistically oriented approach to defining AH and structuring a corpus can be adopted for the purposes of this study.

Such an approach is best demonstrated by Durrant (2009), who, for the purposes of compiling and structuring a corpus, draws disciplinary boundaries based on empirical linguistic evidence. He first compiles and structures a preliminary corpus using the disciplinary map of the University of Nottingham as an initial approximation. For each discipline represented in his preliminary corpus, he produces a list of 'keywords'. Keywords is a corpus linguistics term denoting words which occur more frequently than would be expected by chance in a focus corpus when compared with a reference corpus (Scott, 1999:72). That is to say, keywords reflect the domain of a focus corpus very well. Durrant (2009), therefore, uses his keyword lists to explore similarities in terms of vocabulary use between the disciplines represented in his preliminary corpus. In doing so, he is able to establish which disciplines are most strongly linked in terms of vocabulary use, and thus which share a sense of discourse community. With this information, he restructures his preliminary corpus to produce a linguistically-oriented empirically-derived final corpus. This approach, which will be adopted for the purposes of this study, involves several processes that will be described in more detail in the following subsections: compiling a preliminary corpus (3.1.1), performing a keyword analysis (3.1.2); and restructuring the preliminary corpus (3.1.3).

3.1.1. Compiling a preliminary corpus based on a university disciplinary map

Firstly, then, a preliminary AH corpus was compiled based on the structure of the University of Warwick's Faculty of Arts. The Faculty of Arts, self-described as 'one of the world's top 50 Arts and Humanities faculties' (University of Warwick, 2018), subsumes 7 departments: English and Comparative Literary Studies; Classics and Ancient History; Film and Television Studies; History (including Comparative American Studies); History of Art; Modern Languages and Cultures; and Theatre & Performance Studies and Cultural & Media Policy Studies. This departmental structure is the primary organisational feature of the University of Warwick's research repository (WRAP), in which research papers by the staff and students are electronically stored and made available to download on an open-access basis. WRAP was therefore used as the sole source of corpus data collection in the present study because it is obviously practically advantageous to collect corpus data from one source where it is already conveniently organised. Furthermore, it is conceptually advantageous as it guarantees a certain level of institutional homogeneity in terms of discursal practices and conventions.

It was decided that PhD theses would serve as the fabric of the corpus because of their genre and specificity. As Swales (2004) suggests, PhD theses are strongly oriented towards the

research world (ibid:99) and heavily impacted by discipline specific conventions and expectations (ibid:103). Their positioning in the research-writing genre and discipline-specificity make them not only a suitable source from which to draw implications for EAP, but particularly ESAP. Accordingly, 16 theses from six departments in the Faculty of Arts (96 in total) were collected for the provisional corpus⁵. It is, however, known that disciplinary and institutional conventions and expectations can and do change over time as practitioners orientate to research topics that are recognized as current and relevant. As Foucault (1981:66) states, 'for there to be a discipline, there must be the possibility of formulating new propositions ad infinitum'. Therefore, in order to ensure that the present study is both current and relevant, every effort was made to collect only PhD theses that were completed in the previous five academic years (2013- 2017 inclusive). In certain departments, though, this was not possible as there was a shortage of recently completed PhD theses. This resulted in a final date range of 2013-2017 (81.25%), 2008-2012 (11.45%) and 2003-2007 (7.3%). No theses more than 15 years old were collected.

The 96 theses from six departments between the years of 2003-2017 were cleaned and compiled into a preliminary corpus. The cleaning process required manually removing everything before the introduction (e.g. contents, forewords, abstracts) and everything after the conclusion (e.g. acknowledgements, references and appendices). Following this, the preliminary corpus was compiled using the Sketch Engine, a multifunctional web-based concordancer used extensively in lexicography (Thomas, 2017:6). This produced a preliminary AH corpus comprising approximately 8.7 million words. This corpus was then divided into six subcorpora to represent the six departments of the Faculty of Arts: English and Comparative Literary Studies (ENGCOMP); Film and Television Studies (FILMTV); History (including Comparative American Studies) (HIST); History of Art (ARTHIST); Theatre & Performance Studies and Cultural & Media Policy Studies (THEATRE); and Modern Languages and Cultures (MODLANG) (Table 4). The subcorpora, which each contain an equal number of texts, are all of *slightly* unequal word counts due to the 80,000 word limit of a PhD thesis being more strictly adhered to in some departments than others and the allowance of +/-10%.

Table 4. Provisional corpus based on The University of Warwick's Faculty of Arts⁶

⁵ The department of Ancient History and Classics was excluded from the preliminary corpus due to a shortage of PhD theses in the Faculty of Arts WRAP.

⁶ The Sketch Engine reports the total word count of the corpus as 8,769,964, however it only reports the word counts of subcorpora as approximations (as indicated by the ~). If the approximate word counts for each of the subcorpora are summed, the total word count of the corpus is actually 8,769,960 (4 less than the total reported by the Sketch Engine).

Subcorpus	Words	PhD Theses	Theses Date range
ARTHIST	~1,586,559	16	2003-2017
ENGCOMP	~1,509,793	16	2013-2017
FILMTV	~1,526,945	16	2008-2017
HIST	~1,595,517	16	2013-2017
MODLANG	~1,372,451	16	2010-2016
THEATRE	~1,178,695	16	2003-2017
Total	8,769,964	96	2003-2017

3.1.2. Performing a keyword analysis

With the preliminary AH corpus compiled, the next stage was to analyse the degree of overlap in terms of keyword usage between the subcorpora, that is, the degree of commonality in terms of vocabulary use between departments. The results from this analysis would determine which departments, if any, share a sense of discourse community. In order to perform the keyword analysis, six lists of keywords, one for each subcorpus (department), were produced using the Sketch Engine's Word List tool. For the purposes of the present study, keywords were defined using an adapted version of Durrant's (2009:160) four criteria (Fig. 6).

Fig. 6. Adapted version of Durrant's keyword criteria

In the present study, keywords are lemmas which:

- (1) contain four or more letters
- (2) occur in the subcorpus with a minimum normed frequency of 20 per million words
- (3) occur in at least 25% of texts in the subcorpus
- (4) occur more frequently in the subcorpus than in the British National Corpus (BNC) with an 'add-n' threshold for keyness of 0.1.

It was necessary to adapt Durrant's (2009) criteria to suit the purposes of the present study for two reasons. First, throughout Durrant's study 'words' are analysed, while throughout this study 'lemmas' are analysed⁷. Second, whereas Wordsmith Tools (the corpus analysis software

⁷ See 3.2.1. for a detailed explanation of why 'lemma' is used as the search attribute throughout this study.

used by Durrant) calculates 'keyness' using Log-Likelihood, the Sketch Engine's Word List tool calculates 'keyness' using Simple Math⁸ (Screenshot 1).

Screenshot 1. Producing a list of keywords using the Sketch Engine's Word List tool

Word list options

Subcorpus: FILMTV [Info](#) [create new](#)

Search attribute: lemma (lowercase)

use n-grams. Value of n: from 2 to 2

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression: [a-z]{4,100} [Info](#) **Criterion 1**

Minimum frequency: 30 **Criterion 2**

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: [Choose File](#) No file chosen [Clear](#)

Blacklist: [Choose File](#) No file chosen [Clear](#) [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple

Keywords

Reference (sub)corpus: British National Corpus (BNC) [Info](#) **Criterion 4**

Prefer: rare words common words: 0.1

Change output attribute(s)

--- --- ---

You can select one or more output attributes. Please note that this option can be time-consuming.

[Make word list](#)

Using the aforementioned four criteria in each of the six subcorpora, the Sketch Engine's Word List tool produced six lists of keywords, each of varying lengths between 1300 to 1800 items. In order to compare the lists for overlap, each list needed to be of equal length. Therefore, the lists were ordered by raw frequency (highest to lowest) and shortened to 1000 keywords (that is, the 1000 most frequent keywords in each department).

The keyword lists were ordered by raw frequency rather than the Sketch Engine's keyness score because it was clear that the keywords with the highest keyness score on each list were highly idiosyncratic and thus unlikely to be common to other keyword lists. For example, out of the top 10 keywords with the highest keyness score in ENGCOMP and MODLANG, there is only one commonality⁹ (Table 5), whereas out of the top 10 most frequent keywords in

⁸ Keyness, Simple Math and the 'add-n' parameter are explained in more detail in section 3.2.4 where it is of most significance to the methodology.

⁹ The only commonality between the top 10 most key keywords of ENGCOMP and MODLANG appears to be a non-word (*ofthe*).

ENGCOMP and MODLANG, eight are common to both lists (Table 6). Therefore, because the aim of the keyword analysis was to examine shared discursal practices amongst the departments, raw frequency was a more suitable metric for ranking and shortening the keyword lists. With a list of the top 1000 most frequent keywords for each subcorpus, the degree of commonality in terms of vocabulary use between the departments could be examined.

Table 5. Top ten keywords by keyness score

Top 10 ENGCOMP keywords by keyness score (high to low)	Keyness score	Top 10 MODLANG keywords by keyness score (high to low)	Keyness score
equivocation	215.8	della	285.1
manga	191.4	impegno	213.3
sonnet	177	delle	161.9
allegory	154.3	dell	145.5
allegorical	149.1	narrator	132.1
<u>ofthe</u>	96.4	mafia	128.6
ibidem	74.8	<u>ofthe</u>	126
tran	62.8	letteratura	122.1
trope	60.7	storia	119.1
inthe	60.2	vita	115.4
emigration	50.9	degli	111.2

Table 6. Top 10 keywords by raw frequency

Top 10 ENGCOMP keywords by raw frequency (high to low)	Raw freq.	Top 10 MODLANG keywords by raw frequency (high to low)	Raw freq.
that	19,687	<u>this</u>	8,832
<u>this</u>	9,487	<u>page</u>	7,527
<u>which</u>	7,241	<u>which</u>	7,346
<u>page</u>	5,696	<u>also</u>	3,106
<u>also</u>	3,032	text	2,789
<u>these</u>	2,631	<u>work</u>	2,591
<u>between</u>	2,603	<u>between</u>	2,507
only	2,496	<u>these</u>	2,479
<u>work</u>	2,452	film	2,425
<u>such</u>	2,380	<u>such</u>	2,224

To quantify the commonalities in vocabulary use between each subcorpora, each keyword list was compared to another list and the duplicate values were counted. In other words, two keyword lists were compared and the number of keywords that were common to both lists was summed. For example, of the 1000 most frequent ENGCOMP keywords, 661 of them were common to the 1000 most frequent FILMTV keywords (and vice versa). This process of comparing one list to another was repeated with all six keyword lists to generate a matrix

detailing the degree to which each department within the Faculty of Arts resembles every other in terms of vocabulary use (Table 7).

Table 7. The degree of commonality across departments in terms of vocabulary use

	ENGCOMP	FILMTV	HIST	ARTHIST	THEATRE	MODLANG	Average
ENGCOMP	-	66.10%	51.80%	46.80%	63.90%	64.10%	58.54%
FILMTV	66.10%	-	498	45.50%	63.40%	61.30%	57.22%
HIST	51.80%	49.80%	-	43.20%	51.60%	48.90%	<u>49.06%</u>
ARTHIST	46.80%	45.50%	43.20%	-	45.30%	47.10%	<u>45.58%</u>
THEATRE	63.90%	63.40%	51.60%	45.30%	-	59.00%	56.64%
MODLANG	64.10%	61.30%	48.90%	47.10%	59.00%	-	56.08%
Average	58.54%	57.22%	<u>49.06%</u>	<u>45.58%</u>	56.64%	56.08%	-

As can be seen in the matrix data, four of the subcorpora are strikingly similar in terms of vocabulary use. English and Comparative Literary Studies; Film and Television Studies; Theatre & Performance Studies and Cultural & Media Policy Studies; and Modern Languages and Cultures are all in the 56-59% average commonality bracket, i.e. each of these departments has an average of 560-590 keywords in common with the other five departments. Whereas, History (including Comparative American Studies) and History of Art are clearly less similar in terms of vocabulary use, both falling in the 45-50% average commonality bracket, i.e. 450-500 keywords in common with the other departments. It seems, then, there is a linguistic division within the University of Warwick's Faculty of Arts, with four departments being much more strongly linked in terms of vocabulary use.

3.1.3. Restructuring the preliminary corpus to produce the final corpus

The results from the keywords analysis provide empirical linguistic evidence to suggest that the vocabulary needs of students in the University of Warwick's Faculty of Arts would be better served by *at least* two separate lists of academic collocations. However, this study, which can only realistically aim to produce one list of academic collocations, will focus exclusively on the larger grouping that emerged from the keyword analysis. Therefore, the Academic Collocation List for Arts and Humanities will be derived from a final corpus comprising 5.5 million words from 64 PhD theses across four fields of AH: English and Comparative Literary Studies (ENGCOMP); Film and Television Studies (FILMTV); Modern Languages and Cultures (MODLANG); and Theatre & Performance Studies and Cultural & Media Policy Studies

(THEATRE). Hereafter, this empirically-derived corpus will be referred to as the Arts and Humanities Corpus (AHC) (Table 8).

Table 8. The Arts and Humanities Corpus (AHC) (and its subcorpora)

Subcorpus	Words	PhD Theses	Theses Date range
ENGCOMP	~1,509,793	16	2013-2017
FILMTV	~1,526,945	16	2008-2017
MODLANG	~1,372,451	16	2010-2016
THEATRE	~1,178,695	16	2003-2017
AHC (total)	5,587,887	64	2003-2017

3.2. Computational analysis of the AHC to produce a preliminary list of academic collocations

At this stage, a computational analysis of the AHC was performed using the Sketch Engine's Word List tool to generate a preliminary list of academic collocations. For the purposes of this analysis, it was necessary to further define 'academic collocations' as sets of two lemmas that co-occur within a span of +/-5, co-occurring at least 28 times in total in the AHC and, importantly, co-occurring with a higher frequency in the AHC than in a general English corpus with an 'add-n' threshold for keyness of 0.1 (Screenshot 2). Arriving at this definition involved a number of important theoretical and practical decisions which will be further explained in the following subsections: search attribute (see 3.2.1); span (see 3.2.2); frequency (see 3.2.3); and keyness (see 3.2.4).

Screenshot 2. Producing a list of academic collocations using the Sketch Engine

Word list options

Subcorpus: None (whole corpus) [info](#) [create new](#)

Search attribute: collocations **Search attribute**

use n-grams. Value of n: from 2 to 2

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression: _____

Minimum frequency: 28 **Frequency**

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: No file chosen

Blacklist: No file chosen [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple

Keywords

Reference (sub)corpus: British National Corpus (BNC)
 BNC_NON_ACADEMIC

Prefer: rare words common words 0.1

Change output attribute(s)

You can select one or more output attributes. Please note that this option can be time-consuming.

3.2.1. Search attribute: collocations

The Sketch Engine’s Word List tool is ‘a powerful tool capable of generating many types of lists’ (Thomas, 2017:196). In this sense, ‘Word List’ is somewhat of a misnomer because, along with lists of *words*, it can generate lists of *n-grams*, *terms*, *parts of speech* and *collocations*. The present study uses the search attribute ‘collocations’ (Screenshot 3), which produces lists of collocations with each entry represented as a lemma with a suffixed abbreviation of the part of speech (POS). For example, the collocation *cultural difference* is represented as *cultural-j difference-n*, that is, the lemma *cultural* with the POS suffix *-j* for adjective and the lemma *difference* with the POS suffix *-n* for noun. Therefore, using the search attribute ‘collocations’ in the Sketch Engine’s Word List tool means the collocational analysis will be done on lemmas rather than words because the results are lemmatised.

Screenshot 3. Search attribute: collocations

Subcorpus: None (whole corpus) [info](#) [create new](#)

Search attribute: collocations **Search attribute**

Although collocational analyses can be done on words or lemmas, lemmatisation can mask important collocational relationships between the different inflectional forms of a lemma (see Sinclair, 1991; Tognini-Bonelli, 2001; Hoey, 2005). Take, for example, the lemmatised

collocation *originally-a publish-v* which occurs in the AHC exclusively as *originally published*. There are no co-occurrences of *originally* with the word forms *publish*, *publishes* or *publishing*, meaning lemmatisation masks the true collocational relationship which is not between the lemmas *originally* and *publish* but the words *originally* and *published*. For this reason, many collocational analyses are done on words rather than lemmas (e.g. Durrant, 2009; Ackerman and Chen, 2013).

In the context of the present study, though, it would not be practical to treat each inflectional variation as a separate word because it could result in a very long list of collocations of very little use to EAP teaching-learning. High-frequency collocations such as *create-v space-n*, for example, would need to be listed as *create space*, *creates space*, *created space* and *creating space*, because each inflectional form of the lemma *create* (*creates*, *created* and *creating*) frequently collocates with *space* in the AHC. This is why, for the purposes of presenting the ACL, Ackerman and Chen (2013:245), who analysed words rather than lemmas, ultimately listed independent collocations such as *fundamental assumption* and *fundamental assumptions* as one entry under the lemmas *fundamental assumption*. It seems that for certain purposes, such as developing an academic collocation list, working with lemmas is, as Hoey (2005:5) puts it, 'useful'.

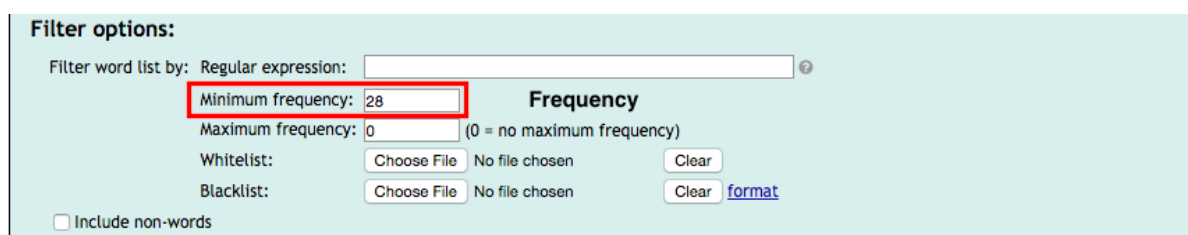
3.2.2. Span: +/-5

For the purposes of the computational analysis, the span within which co-occurring lemmas were considered collocates is +/-5. In collocation analyses, spans of +/-2 (Clear, 1993), +/-3 (Gledhill, 2000; Ackerman and Chen, 2013), +/-4 (Sinclair et al., 2004; Durrant, 2009) and +/-5 (Xu et al., 2003; Stuart and Trelis, 2006) have all been used, but Jones and Sinclair's (1974) determination that the optimum span for identifying significant collocations is +/-4 remains fairly uncontroversial. A span of +/-4, then, would have been the researcher's first choice for the present analysis, however the Word List tool in the Sketch Engine does not allow the span to be adjusted – the default span is fixed at +/-5. Fortunately, though, this is a suitable span for the present study in which one of the key aims is to identify positionally variable collocations, such as **context of cinema** [+/-2] and **context of contemporary mainstream Hollywood cinema** [+/-5].

3.2.3. Frequency: 28

The minimum raw frequency threshold for lemmas co-occurring within a span of +/-5 in the AHC was set at 28 (Screenshot 4). The Sketch Engine allows only raw frequencies to be entered, but because it is typical for collocation studies to operationalise normed frequencies, the raw frequency of 28 was calculated as equivalent to a normed frequency of 5 occurrences per million words ($5 \times 5,587,887 / 1,000,000 = 27.939435$). A normed frequency threshold of 5 p/m words is reasonably high in comparison to similar studies. For example, Ackerman and Chen (2013) and Durrant (2009) both applied a normed frequency threshold of just 1 p/m words. Yet, it was felt by the present researcher that collocations occurring on average of once per 12.5 PhD theses (at 80,000 words per thesis) were not ecologically valid. Furthermore, a lower frequency threshold, such as 1 p/m words, would result in a very large and unmanageable preliminary data set. The use of the higher normed frequency threshold of 5 p/m words, therefore, ensured that the computational analysis would yield a list of collocations that was both ecologically valid and manageable within the constraints of the study.

Screenshot 4. Frequency: 28



The screenshot shows the 'Filter options' section of the Sketch Engine interface. It includes a 'Filter word list by: Regular expression:' field. Below it, the 'Minimum frequency' is set to 28, which is highlighted with a red box. The 'Maximum frequency' is set to 0, with a note '(0 = no maximum frequency)'. There are 'Whitelist:' and 'Blacklist:' sections, each with a 'Choose File' button and a 'Clear' button. A 'format' link is also present. At the bottom, there is a checkbox for 'Include non-words'.

3.2.4. Keyness: add-n threshold of 0.1

In order to identify 'academic' vocabulary, that is, vocabulary which is more frequent in academic than non-academic texts, some list compilers, namely Coxhead (2000) and Ackerman and Chen (2013), follow a process that involves excluding the most frequent 2000 words of English as represented by the outdated GSL (West, 1953). As previously discussed, this can be problematic and therefore the present study adopts a different approach to identifying 'academic' collocations which uses the 'Keywords Output Option' of the Sketch Engine's Word List tool. This method will identify collocations which are 'key' to the AHC, that is, collocations which occur significantly more frequently in the AHC than in a reference corpus of general English texts. The 'Keywords Output Option' calculates 'keyness' using Kilgarriff's (2009) Simple Math for keywords.

Simple Math (Kilgarriff, 2009) takes the normed frequency of a collocation in a focus corpus (FC) and divides it by the normed frequency of the same collocation in a reference corpus (RC) to produce a ratio representing the frequency of the collocation in the FC to the RC (FC:RC ratio). What is unique about Simple Math (and to the Sketch Engine), though, is that the researcher is able to manipulate the FC:RC ratio by ‘adding one’ using the ‘add-n’ function in the ‘Keywords Output Option’. By adding more ‘ones’ to the FC:RC ratio the researcher is able to identify commoner keywords (or collocations) and by adding fewer ‘ones’ to the FC:RC ratio the researcher is able to identify rarer keywords (or collocations). This Simple Math approach to calculating ‘keyness’ is situated in contrast with the so-called sophisticated math approach, which generally uses MI, Log-Likelihood or Fisher’s Exact Test. These sophisticated approaches all need a null hypothesis to build on - that language is random - but this null hypothesis does not exist (as it is known that language is not random), and therefore, according to Kilgarriff (2009:2), there are no theoretical reasons for using sophisticated math over Simple Math to calculate keyness.

For the purposes of this study, then, the FC is the AHC and the RC is the BNC_NON_ACADEMIC, an 80 million-word non-academic subcorpus of the BNC created using Lee’s (2001) classifications. The Simple Math calculation was manipulated using the ‘add-n’ function, adding *just* 0.1 (Screenshot 5). By adding only a fraction of a ‘one’ to the Simple Math calculation, the Sketch Engine will identify collocations with a high FC-to-RC ratio, or, in other words, collocations in the AHC that are particularly *rare* in the BNC_NON_ACADEMIC. For example, *dominant culture* which occurs 111 times in the AHC and 0 times in the BNC_NON_ACADEMIC and *key element* which occurs 71 times and 0 times, respectively. This high FC:RC ratio, which ensures ‘rareness’ in non-academic texts, means that budding EAP students will be less likely to have encountered the collocations the computational analysis produces before their EAP journey begins.

Screenshot 5. Keyness: add-n threshold of 0.1

The screenshot shows the 'Output options' section of the Sketch Engine interface. It includes the following elements:

- Output options:**
 - Frequency figures:** Three radio buttons are present: 'Hit counts' (selected), 'Document counts', and 'ARF'.
 - Output type:** Two radio buttons are present: 'Simple' and 'Keywords' (selected).
- Keywords section (highlighted with a red box):**
 - Reference (sub)corpus:** A dropdown menu showing 'British National Corpus (BNC)' and 'BNC_NON_ACADEMIC' (selected).
 - Prefer:** A slider between 'rare words' and 'common words'. The slider is positioned towards 'rare words', with a numerical value of '0.1' displayed next to it.
- Keyness:** A label on the right side of the interface.

3.2.5. Results from the computational analysis

With the search attribute 'collocation' (which denotes lemmas rather than words), the span of +/-5, the raw frequency of 28 (normed frequency of ~5 p/m words) and the 'add-n' threshold for keyness set to 0.1, the Sketch Engine's Word List tool yielded a preliminary list of academic collocations with 5,222 entries. This list, though, clearly requires refinement if it is to be of any use in EAP teaching-learning (Screenshot 6). The following subsections, therefore, describe the process of manual refinement.

Screenshot 6. Results from the computational analysis

Word list

Corpus: ARTS AND HUMANITIES

Reference corpus: British National Corpus (BNC)
Reference subcorpus: BNC_NON_ACADEMIC
[Switch focus and reference \(sub\)corpus](#)

Page 1 Go [Next >](#)

		ARTS AND HUMANITIES		British National Corpus (BNC) : BNC_NON_ACADEMIC		
ws_collocations		frequency	frequency/mill	frequency	frequency/mill	Score
not-a	verbs modified by "%w" be	5,906	865.1	0	0.0	8651.5
not-a	verbs modified by "%w" do	5,201	761.8	0	0.0	7618.9
also-a	verbs modified by "%w" be	2,439	357.2	0	0.0	3573.4
not-a	adverbs modified by "%w" only	2,029	297.2	0	0.0	2972.9
only-a	modifiers of "%w" not	2,029	297.2	0	0.0	2972.9
as-a	adverbs modified by "%w" well	1,748	256.0	0	0.0	2561.3
well-a	modifiers of "%w" as	1,748	256.0	0	0.0	2561.3
university-n	nouns modified by "%w" press	1,339	196.1	0	0.0	1962.2
new-n	nouns modified by "%w" york	1,162	170.2	0	0.0	1703.0
only-a	verbs modified by "%w" be	1,107	162.1	0	0.0	1622.4
not-a	verbs modified by "%w" have	1,086	159.1	0	0.0	1591.7
hong-n	nouns modified by "%w" kong	889	130.2	0	0.0	1303.1
often-a	verbs modified by "%w" be	744	109.0	0	0.0	1090.7
same-j	nouns modified by "%w" time	737	107.9	0	0.0	1080.5
so-a	verbs modified by "%w" do	699	102.4	0	0.0	1024.8
still-a	verbs modified by "%w" be	648	94.9	0	0.0	950.1
always-a	verbs modified by "%w" be	603	88.3	0	0.0	884.2
here-a	verbs modified by "%w" be	591	86.6	0	0.0	866.6
also-a	verbs modified by "%w" see	586	85.8	0	0.0	859.3
also-a	verbs modified by "%w" have	575	84.2	0	0.0	843.2
work-n	pronominal possessors of "%w" his	562	82.3	0	0.0	824.2
already-a	verbs modified by "%w" have	528	77.3	0	0.0	774.4
case-n	nouns modified by "%w" study	523	76.6	0	0.0	767.0
nineteenth-j	nouns modified by "%w" century	520	76.2	0	0.0	762.6
other-j	nouns modified by "%w" word	508	74.4	0	0.0	745.1
longer-a	modifiers of "%w" no	480	70.3	0	0.0	704.0
no-a	adverbs modified by "%w" longer	480	70.3	0	0.0	704.0

Before this qualitative refinement process begins, though, it is important to note that quantitative dispersion and significance thresholds have not yet been set. Although these thresholds would typically be set during the initial computational analysis detailed above, the Sketch Engine's Word List Tool does not provide these statistics and thus thresholds cannot be set at this stage. Consequently, the range of each individual collocation needs to be

calculated by counting concordance lines¹⁰, and significance scores (e.g. MI, t-score or logDice) need to be retrieved manually for each entry¹¹. Because performing these two tasks for all 5,222 entries would be time-consuming beyond the constraints of the present study, it was decided that the dispersion and significance thresholds would be applied after the list had been manually refined.

3.3. Manual refinement of syntactic combinations

As previously discussed, grammatical collocations, that is, the combination of one open-class and one closed-class component (e.g. *were collected* and *in addition*), consist of relatively fixed patterns that can be more easily internalized into the learner lexicon and thus have not attracted much research attention. The present study, therefore, focuses exclusively on lexical collocations, that is, the combination of two open-class components (e.g. *clearly demonstrate* and *inhabit space*), meaning the syntactic relationship between individual components is an important consideration.

The phraseological approach to collocation, unlike the neo-Firthian approach, consistently requires that the components of a collocation should be syntactically related (Nesselhauf, 2005:17) and consequently there are a number of pre-defined syntactic combinations proposed by phraseologists. While some, such as Aisenstadt (1981) and Benson et al. (1997), allow the syntactic combination of one open-class and one closed-class component (i.e. grammatical collocations), Hausmann (1989:1010) allows only the syntactic combination of two open-class components (i.e. lexical collocations). Therefore, his six syntactic combinations (Fig. 7) were deemed most suitable for the purposes of the present study.

Fig. 7. Hausmann's syntactic combinations

Hausmann's (1989:1010) syntactic combinations:

1. Adjective + noun (e.g. social role)
2. Noun + noun (e.g. stage production)
3. Subject noun + verb (e.g. chapter examine)
4. Verb + object noun (e.g. occupy space)

¹⁰ It is possible to calculate the range of single-word vocabulary in the Sketch Engine's Word List tool using a Whitelist (albeit as part of a somewhat convoluted process), but this process is not possible when working with collocations.

¹¹ The Sketch Engine's Word List tool provides significance scores (logDice) for collocations when working with *only* a focus corpus, but not when working with both a *focus* and *reference* corpus, as in the case of the present study.

5. Adverb + adjective (e.g. mutually exclusive)
6. Verb + adverb (e.g. focus specifically)

The preliminary list with 5,222 entries was exported from the Sketch Engine to Microsoft Excel so that non-target combinations, that is, combinations which do not fit Hausmann's (1989:1010) six pre-defined syntactic combinations, could be manually identified and removed. The identification and removal of non-target combinations was facilitated by the extremely accurate POS suffixes attached to each entry¹². For example, collocations such as *almost-a exclusively-a* (adverb + adverb) and *cultural-j literal-j* (adjective + adjective) were quickly and easily identified as non-target combinations by their POS suffixes and removed.

The POS suffixes, though, only make very basic distinctions between word classes, therefore it was necessary to identify some of the more nuanced parts-of-speech without the use of the suffixes. For instance, certain adverbs can function as conjunctions (e.g. *therefore*, *also* and *thus*), other adverbs can function as qualifiers (e.g. *very*, *too* and *quite*) and *have* can function as an auxiliary verb. These closed-classes (conjunctions, qualifiers and auxiliary verbs) cannot be quickly and easily distinguished by their POS suffixes (e.g. *therefore-a*, *very-a* and *have-v*), therefore in many cases it was necessary to examine concordance lines. Take, for example, the collocation *narrator-n have-v*. The POS suffix *-v* simply indicates verb, yet *have* could be functioning as an auxiliary verb (e.g. *the narrator has ultimately lost control of his metaphor*) or a verb meaning *to possess* (e.g. *the narrator has a split identity*). The former would be a non-target combination as auxiliary verbs are closed-class, while the latter would be a target combination (subject noun + verb) as verbs are open-class. Identifying and distinguishing the subtle differences between these word classes was a time-consuming process.

Once all non-target combinations were removed (Table 9), the list included 3,766 entries.

Table 9. Examples of non-target combinations (one open and one closed class component)¹³

Non-target combination	Collocations
------------------------	--------------

¹² The Penn TreeTagger, the Sketch Engine's default tagset for labelling the POS of each token in a corpus, boasts an extremely high-level of tagging accuracy (up to 98%) (Sketch Engine, 2018).

¹³ This list is not intended to be exhaustive – in many cases, non-target combinations were composites of two closed-class components e.g. *however be* (conjunction + auxiliary verb) or *really have* (qualifier + auxiliary verb).

conjunction + verb	e.g. <i>also acknowledge, as discuss</i>
preposition + noun	e.g. <i>between space, amongst women</i>
pronoun + noun	e.g. <i>our society, her article</i>
auxiliary verb + adverb	e.g. <i>be deeply, have far</i>
qualifier + adjective	e.g. <i>very clear, quite different</i>
verb + question word	e.g. <i>explore how, highlight how</i>

3.4. Manual removal of noise

The purpose of this stage was to further refine the list by removing noise. It was decided for various practical and theoretical reasons that entries fitting twelve descriptions would be considered noise and removed (Fig. 8).

Fig. 8. Noise removal criteria

<p>Entries fitting the following descriptions were considered noise:</p> <ol style="list-style-type: none"> 1. Collocations containing non-words 2. Collocations containing non-English words 3. Collocations containing proper nouns 4. Collocations containing the noun 'page' 5. Collocations containing concrete geographical and seasonal references 6. Collocations containing adverbs of frequency 7. One collocation from a pair of duplicates 8. Collocations containing cardinal or ordinal numbers 9. Linguistically incomplete units 10. Collocations containing partial or full titles of sources 11. Collocations with two identical components 12. Collocations containing the adverb 'not'
--

It is, though, important to explain why these criteria were applied. While criteria 1 and 2 require little explanation, criteria 3 through 9 were developed in response to qualitatively examining the remaining entries on the list and closely considering the decisions of other researchers who have compiled lists of collocations for EAP. Durrant (2009:162) excludes proper nouns (3); items which occurred chiefly outside the main text (4); items which appear on the listing twice (7); and numbers (8). Ackerman and Chen (2013: 239) exclude concrete geographical references (5); adverbs of frequency (6); and linguistically incomplete units (9). It was clear from examining the present list that collocations fitting Durrant's (2009) and Ackerman and Chen's (2013) criteria constituted a very large proportion of entries yet offered very little in terms of pedagogical value. For example, duplicates (7), which represent the multidirectional relationships of collocations (e.g. *citizenship* as a collocate of *multicultural* and *multicultural* as a collocate of *citizenship*), constituted almost half of the remaining entries on the present list.

In the case of duplicates, only the collocation in the syntactic order discussed above (see 3.3) was retained (e.g. *multicultural citizenship* [adjective + noun])¹⁴. Further examples which demonstrate why it was necessary to apply criteria 1-9 can be found in Table 10.

Table 10. Collocations which were removed as noise (criteria 1-9)

Criteria no.	Description	Examples
1	Collocations containing non-words	<i>chapter m√°t√©, √©migr√© novel</i>
2	Collocations containing non-English words	<i>sans fleurs, letteratura italiana</i>
3	Collocations containing proper nouns	<i>early England, Hollywood film</i>
4	Collocations containing the noun 'page'	<i>page ii</i>
5	Collocations containing concrete geographical and seasonal references	<i>Chinese nationalism, early summer</i>
6	Collocations containing adverbs of frequency	<i>often use, never see</i>
7	One collocation from a pair of duplicates	<i>multicultural citizenship, citizenship multicultural</i>
8	Collocations containing cardinal or ordinal numbers	<i>chapter three, twentieth century</i>
9	Linguistically incomplete units	<i>class people as in working class people, golden film as in golden age film</i>

Although the aforementioned nine criteria removed much of the noise from the list, it was clear from the remaining entries that three more criteria would be needed to suitably remove all noise. These criteria, criteria 10 through 12, were the removal of titles of sources that appeared in the main text (10), collocations with two identical components which were chiefly the result of listing or comparing (11)¹⁵ and the adverb *not* which was simply used to express the negative form of a verb or adjective (12). These criteria were not formulated in consideration of existing literature, but rather in consideration of the remaining entries on the list (Table 11).

Table 11. Collocations which were removed as noise (criteria 10-12)

Criteria no.	Description	Examples
--------------	-------------	----------

¹⁴ In the case of adverb + verb and noun + noun combinations, it was necessary to check concordance lines to ascertain in which direction the collocation was most common (e.g. *focus specifically* or *specifically focus* and *time and space* or *space and time*) – only the most common co-occurrence pattern was retained.

¹⁵ Entries such as *culture-n culture-n* were the result of concordances such as '*Chinese **culture**, Japanese **culture** and modern American **culture***' and '*the textual comparisons reveal differences in the source **culture** and target **culture***'.

10	Collocations containing partial or full titles of sources	<i>Divine Comedy, Paradise Lost</i>
11	Collocations with two identical components	<i>culture culture</i>
12	Collocations containing the adverb 'not'	<i>not know, not see, not possible</i>

It is important to note that two further criteria, both of which were applied by Ackerman and Chen (2013), were also originally applied to the present list as part of the noise removal process. However, upon reflection, the researcher concluded that the removal of collocations based on these criteria should be reversed. These criteria were:

13. The removal of adverbs of time (e.g. *already mention* and *previously discuss*)
14. The removal of common transparent adjectives (e.g. *good example*)

Criteria 13 removed collocations such as *soon become* and *still hold*, yet it also removed those such as *previously discuss* and *already mention*. The former are arguably of little pedagogical relevance, however the same could not be so easily argued for the latter. In fact, although Ackerman and Chen (2013) claim to have removed adverbs of time from their list, there are five entries on the final ACL containing the adverb of time *previously* (*previously described, previously discussed, previously known, previously mentioned* and *previously thought*). Therefore, it was decided that in the spirit of consistency all target combinations containing an adverb of time should remain on the list (in the hope that accretions such as *soon become* and *still hold* might be removed by a later stage of manual refinement).

Criteria 14 was initially used to remove collocations containing the adjective *good*, such as *good example* and *good way*. This is because *good* is the only example of a 'common transparent adjective' cited by Ackerman and Chen (2013:239). Yet, they are not explicit in describing what constitutes 'common' and/ or 'transparent' in their study. For this reason, it was difficult for the researcher in the present study to draw a clear boundary between common/ uncommon and transparent/ non-transparent adjectives. For example, the adjectives *black* and *white* might be considered transparent when used to describe a *car* or a *plate*, but perhaps not when used to describe a *character*. Therefore, rather than justify individual decisions regarding the transparency of every adjective on the list, the researcher decided to reverse this criterion and restore all adjectives, including *good* (again in the hope that accretions such as *good example* might be removed by a later stage of manual refinement).

Once the noise removal process was complete, there were 1,101 entries remaining on the list.

3.5. Computational retrieval of dispersion values

At this point, with a much-reduced list of collocations, the dispersion values were retrieved from the Sketch Engine. First, though, the dispersion value threshold was set at a minimum of one occurrence per subcorpus of the AHC, that is to say, each entry had to occur at least once in each of the four fields of AH to remain on the list. In order to compensate for the slightly different sizes of each subcorpus, it would have been preferable to use normed frequencies to calculate dispersion (as was done with frequency), yet due to the relatively small sizes of the subcorpora, using normed frequencies would not have had a compensatory effect. For example, a normed frequency of 1 p/m words would be equivalent to a raw frequency of 1.5 in the 1.5 million-word FILMTV subcorpus and 1.2 in the 1.2 million-word THEAT subcorpus, but a word cannot occur 0.5 or 0.2 times. Therefore, the range threshold was set as a raw frequency of one occurrence per subcorpus which, for the purposes of comparison, is equivalent to a normed frequency of ~0.65 p/m words in the smallest subcorpus (THEAT) to ~0.84 p/m words in the largest subcorpus (FILMTV). These normed frequencies are comparable to those of similar studies, e.g. Durrant's (2009) 1 occurrence p/m words in each discipline and Ackerman and Chen's (2013) 0.2 occurrences p/m words in each discipline.

Because the dispersion value threshold cannot be set in an automated computational analysis in the Sketch Engine, the researcher had to examine and count the concordance lines for each of the 1,101 remaining entries. For example, as can be seen in Screenshot 7, *key factor* occurs in ENGCAMP 1 time, FILMTV 11 times, MODLANG 7 times and THEATRE 9 times. *Key factor*, therefore, remained on the list and the dispersion values were recorded in Microsoft Excel. If an entry did not occur in one of the subcorpora, it was removed from the list.

In all, this process removed 492 entries (Table 12), leaving 609 entries on the list.

Screenshot 7. Counting concordance lines to record dispersion values

Query 28 (4.10 per million) ⓘ

1	ENGCOMP_PH...	decry emigration as a key factor in high celibacy rates among Ireland's
2	FILMTV_PHD...	of the reasoning behind the selection. A key factor determining the selection is the need to test
3	FILMTV_PHD...	of history is created through three key factors :) Period authenticity (historical
4	FILMTV_PHD...	equal happiness, but positive change is a key factor for accomplishing a higher plane of
5	FILMTV_PHD...	of history that is most fascinating, a key factor at the heart of new historical cinema that
6	FILMTV_PHD...	The Drum was a British production was the key factor in the reaction it inspired. It could be taken as
7	FILMTV_PHD...	question of adaptation is consequently a key factor in my analysis of my key film texts. My
8	FILMTV_PHD...	'87 and cites his 'inbetweenness' as a key factor for the creation of his ambiguous persona. As
9	FILMTV_PHD...	climate is commonly considered a key factor that facilitated the twenty-first century
10	FILMTV_PHD...	Johnson. Besides World War II, a key contextual factor that has been attributed to superheroes'
11	FILMTV_PHD...	argues that digital effects are one of the key factors that facilitated the new-wave of superhero
12	FILMTV_PHD...	return numerous times, at one point being key factor contributing a period in which Tony lost his
13	MODLANG_PH...	for a period of two centuries, became a key factor in its fall from prominence. 67 Based on the
14	MODLANG_PH...	marvellous nature - was to become one of the key factors in its astonishing success as a popular printed
15	MODLANG_PH...	novels and plays far from innocent - a key factor in Monénembo's re-casting of African subjects
16	MODLANG_PH...	it adheres to its norms and poetics, will be a key factor in the analysis of the translated play texts.
17	MODLANG_PH...	, which is regulated by the mentioned key factors of power, patronage, ideology and poetics. The
18	MODLANG_PH...	. The chronological element as one of the key factors in translation in general and in theatre
19	MODLANG_PH...	that having emigrated to Britain was a key factor for their originality and modernity. Indeed,
20	THEAT_PHD...	and the constraints of legislation. Key factors of the conditions under which the companies
21	THEAT_PHD...	support is clearly in evidence, was a key factor in achieving the challenging repertory of the
22	THEAT_PHD...	which it signified, which was the key factor in driving this aspect of the development of
23	THEAT_PHD...	devices for individual experience and key factors in decision-making. Bodies of Crisis:
24	THEAT_PHD...	, [t]he trade of women on display was a key factor in the critical success of the Folies, and while
25	THEAT_PHD...	appearance of local actors is another key factor for Mattani in her adapted versions. Many Asian
26	THEAT_PHD...	citizenship, cultural rights have become a key factor in multicultural citizenship. Thirdly, the
27	THEAT_PHD...	shared culture, which are the key internal factors of their identity. However, external factors
28	THEAT_PHD...	ability to speak the Hakka language became a key factor in recognising the Hakka ethnic status. Many

Table 12. Examples of collocations removed by dispersion value threshold

Collocation	ENCOMP	FILMTV	MODLANG	THEAT	Raw Freq.
<i>focus-n group-n</i>	0	0	0	50	50
<i>cultural-j policy-n</i>	3	0	1	332	336
<i>personal-j interview-n</i>	135	133	0	2	270
<i>adopt-v genre-n</i>	0	31	0	0	31
<i>research-n participant-n</i>	0	1	0	33	34

3.6. Manual refinement based on degree of fixedness

As previously discussed, phraseology literature is largely concerned with the manual classification of collocations based on their varying degrees of fixedness. Open and restricted collocations are subject to varying degrees of arbitrary restriction which make them more challenging for learners than idioms and other highly-fixed combinations (Ackerman and Chen, 2013:236). The present study, therefore, focuses exclusively on open and restricted collocations rather than highly-fixed collocations. Yet, it would be beyond the constraints of this study to make and justify individual decisions regarding the degree of fixedness of each entry on the list, which, as a matter of judgement, cannot be established beyond doubt (Howarth, 1996:41). Instead, each remaining entry was cross-checked in the Academic Formulas List (AFL) (Simpson Vlach and Ellis, 2010) and the Longman Dictionary of

Contemporary English Online (LDOCE). If they were listed independently by either of these sources, they were deemed to be highly-fixed and removed. For example, the entry *same-j way-n* was removed because the concordance lines revealed it occurs predominantly as part of the AFL formula *the same way as*, and the entry *case-n study-n* was removed because it is listed as a countable noun in the LDOCE. Open collocations (e.g. *clearly show*) and restricted collocations (e.g. *shoot [a] film*) do not feature in the AFL or LDOCE, and therefore remained on the list.

In total, this process removed 54 entries (Table 13), leaving a total of 555 on the list.

Table 13. Examples of collocations with a high degree of fixedness

Collocation	Source
<i>vantage-j point-n</i>	LDOCE as 'vantage point' (noun, countable)
<i>same-j time-n</i>	AFL as 'at the same time'
<i>wide-j range-j</i>	AFL as 'a wide range'
<i>father-n figure-n</i>	LDOCE as 'father figure' (noun, countable)
<i>important-a role-n</i>	AFL as 'important role in'

3.7. Computational retrieval of statistical significance values

At this stage, with a list of 555 collocations, statistical significance values, that is, values which indicate the strength of association between the components of a collocation, were retrieved from the Sketch Engine using the Word Sketch tool. A Word Sketch is a one-page summary of a words grammatical and collocational behaviour which is presented as a table of collocates with each column representing a different grammatical relationship (Thomas, 2017:176). For example, a Word Sketch for the adjective *new* provides a table of collocates with one column representing nouns modified by *new*. For each collocate in the table, there is a statistical significance value to indicate the strength of association between the node (e.g. *new*) and the collocate (e.g. *perspective*) (Screenshot 8). The statistical significance value which is provided by a Word Sketch is a logDice score.

Screenshot 8. Word Sketch for 'new' (nouns modified by 'new' and logDice scores)

new (adjective) Alternative PoS: noun (freq: 3,201)
 ARTS AND HUMANITIES freq = 5,641 (826.23 per million)

modifiers of "new"	2.55	nouns modified by "new"	89.61	"new" and/or ...	19.59	prepositional phrases		
relatively	19	form +	175	national	40	"new" to ...	15	0.27
is relatively new	10.33	new forms of	9.39	a new national identity	9.68	"new" in ...	12	0.21
radically	7	way +	145	old	28	verbs complemented by "new"		
radically new regimes of knowledge	9.68	new ways of	9.25	old and new	9.39	0.55		
entirely	18	technology	93	social	39	create	4	11.06
an entirely new	9.62	new technologies	9.07	the new social	8.87	create something new .		
something	4	identity	89	critical	20	be	10	8.78
be something new	9.61	new national identity	8.51	new critical	8.73	is nothing new		
nothing	4	generation	61	italian	20	verbs before "new"		
is nothing new	9.53	a new generation of	8.51	new Italian	8.54	1.40		
totally	6	possibility	52	digital	13	be	54	5.95
not totally new	9.31	new possibilities	8.29	new digital technologies	8.43	is not new		
completely	12	meaning	58	urban	13	subjects of "be new"		
a completely new	9.11	new meaning	8.28	the new urban	8.31	1.65		
wholly	5	africa	49	cultural	28	something	29	10.92
a wholly new	8.96	in the new South Africa	8.13	a new cultural	8.27	something new		
not	14	kind	45	political	25	nothing	9	10.58
is not new	3.95	a new kind of	8.05	a new political	8.27	is nothing new		
also	4	medium	44	literary	17			
also new	3.46	new media	7.94	new literary	8.19			
		perspective	43	many	16			
		new perspectives	7.86	many new	8.10			
		nation	38	creative	11			
		of the new nation	7.76	WAC's new Creative Space	8.03			
		mode	37	modern	15			
		new modes of	7.70	a new , modern	7.97			
		idea	38	south	10			
		new ideas	7.67	new South African identities	7.94			
		space	54	theatrical	10			
			7.65		7.90			

Unlike the more commonly used significance measures of MI or t-score, logDice operates on a standardised scale (0-14), does not suffer low-frequency bias (two exclusively associated collocations would both receive a logDice score of 14 regardless of their respective frequencies) and is not based on the assumption that language is random (expected frequency is not included in its equation) (Gabaslova, Brezina and McEney, 2017:165). For these reasons, logDice is arguably a more reliable measure than MI or t-score which are 'largely used as apparently effective, but not fully understood mathematical procedures' (Gabaslova et al., 2017:161). Therefore, logDice scores for all remaining 555 entries were retrieved using the Word Sketch tool and recorded in Microsoft Excel.

It is important to note that no minimum value threshold for logDice score was set. It was observed that the prior stages of computational analysis and manual refinement had cleared the list of collocations in which the association between components was weak (Table 14). For example, of the full band of logDice scores across the 555 entries, the lowest was 6.6 (*new culture*) which, on a standardised scale of 0-14, could be considered to indicate an association of moderate strength. Therefore, rather than set an arbitrary value threshold, it was decided that the expert review would be used to explore whether an empirically derived value threshold could be set.

Table 14. Band of logDice scores (high, medium and low) of the full list (555 items)

Collocation	logDice score	logDice band
-------------	---------------	--------------

mutually-a exclusive-j	13.25 (highest)	high
audience-n member-n	12.54	
stark-j contrast-n	12.35	
encourage-v audience-n	9.18	medium
key-j factor-n	9.17	
central-j concern-n	9.17	
new-j language-n	7.12	low
different-j space-n	6.99	
new-j culture-n	6.6 (lowest)	

3.8. Manual refinement by expert review

The final stage of manual refinement was an intuition-based review of the list by a panel of EAP experts. These experts were the final arbiters of pedagogical relevance and it was hoped that their judgements would provide the means by which to both weed the list of certain entries and prioritise remaining entries. The process of manual refinement by a panel of EAP experts will be outlined in the following subsections: preparing the expert review (3.7.1); and analysing the results from the review (3.7.2).

3.8.3. Preparing the expert review

As previously discussed, Simpson-Vlach and Ellis (2010) and Ackerman and Chen (2013), the respective compilers of the AFL and the ACL, both used expert review in the process of compiling their lists. While Ackerman and Chen (2013) asked a panel of six experts from different professional backgrounds to rate all 4,558 entries on their list, Simpson-Vlach and Ellis (2010) asked a panel of twenty EAP practitioners to rate a representative subset of 108 formulas. Because it would be beyond the constraints of a 4-month unfunded project to find 20 participants *or* to ask participants to rate the full list of 555 entries, the present study amalgamates the two aforementioned approaches. Thus, a panel of five EAP experts from the University of Warwick's Centre of Applied Linguistics and Faculty of Arts (Fig. 9) were asked to review a representative subset of collocations. The data from the expert review of the subset would subsequently be used as the basis for correlation analyses in order to generalise the findings to the full list.

Fig. 9. The expert panel

Expert 1: Principal Teaching Fellow in Applied Linguistics and EAP practitioner
 Expert 2: Senior Teaching Fellow in Applied Linguistics and EAP practitioner
 Expert 3: Teaching Fellow in Applied Linguistics and ESAP practitioner
 Expert 4: Associate Teaching Fellow in the School of Modern Languages and EAP practitioner
 Expert 5: Senior Teaching Fellow in Applied Linguistics and EAP director of studies

If the findings from the subset were to be generalised to the full list, the subset must be broadly representative of the full list. Simpson-Vlach and Ellis (2010:496), who also used correlation analysis to generalise findings from their subset to their full list, ensured that their subset was representative of their full list on three factors: MI score; frequency; and formula length. In light of this, the present subset was made to represent the full list on three factors: (1) logDice score band (high, medium and low), (2) raw frequency band (high, medium and low) and (3) syntactic combinations (all six syntactic combinations). Producing this representative subset was achieved by ordering the remaining 555 entries by their logDice score (from largest to smallest value) in Microsoft Excel and then extracting every fifth entry for the subset. This produced a subset of 112 collocations which was definitely representative of the full list in terms of logDice score, yet not necessarily in terms of raw frequency or syntactic combination. Thus, to check whether the subset was representative of the full list in terms of these two further factors, the subset and full list were compared (Table 15). The results of this comparison show that the subset is representative of the full list on all three factors: (1) logDice, (2) raw frequency and (3) syntactic combination.

Table 15. Comparison: Full list (555 items) versus subset (112 items)

	Full list (555 items)	Subset (112 items)
LogDice band (Average logDice score)	13.25 - 6.6 (9.27)	13.25 - 7.19 (9.3)
Raw Frequency band (Average raw frequency)	366 – 28 (54.5)	300 – 28 (55.3)
Syntactic Combination:		
Adjective + noun	383 (69%)	82 (73.2%)
Noun + noun	68 (12.3%)	10 (8.9%)
Subject noun + verb	8 (1.4%)	2 (1.8%)
Verb + object noun	41 (7.4%)	8 (7.1%)
Adverb + adjective	12 (2.2%)	2 (1.8%)
Verb + adverb	43 (7.7%)	8 (7.1%)

With the representative subset selected, the 112 collocations were prepared for expert review. It was decided that the experts should meet the collocations in context, which necessitated retrieving a concordance line which exemplified a 'typical' occurrence for all 112 collocations. For example, in the collocation *film-n show-v*, the node *film* most frequently occurs in its base form, while the collocate *show* most frequently occurs with the third person verbal inflection *-s* in span position +1. Therefore, the collocation *film-v show-v* was presented to the experts in its most 'typical' occurrence as *the **film shows** a less sympathetic view of the IRA if examined carefully*. All 112 collocations were represented in context and presented to the experts with an overview of the study (Appendix 1). The experts were asked to consider two questions and then use a Likert scale to give each collocation a score of 1-4 (Fig. 10).

Fig. 10. Questions and Likert scale for expert review

Please review each collocation in consideration of these two questions:

- Is it appropriate to consider the entry as an academic collocation for ESAP teaching-learning purposes?
- Do you think the collocation is worth teaching explicitly as part of an Arts and Humanities ESAP course?

Once you have considered these two questions please give each collocation **one** score between 1 and 4 based on the following scale:

- definitely exclude
- perhaps exclude
- perhaps include
- definitely include

The questions were made intentionally vague to allow for the individual perspectives of experts from heterogenous backgrounds (Applied Linguistics and Arts and Humanities). They were also non-technical to account for the multitude of definitions and approaches to collocation that exist. It is, after all, impossible to know with which of the plethora of collocation literature the experts will be most conversant (e.g. neo-Firthian and/ or phraseological).

3.8.2. Results from the expert review

The results from the expert review were collated in Microsoft Excel and the Intraclass Correlation Coefficient (ICC) was calculated to describe the level of agreement between the five experts. The Intraclass Correlation Coefficient (ICC) was 0.315, which, according to Cicchetti's (1994) oft-cited guidelines, denotes 'poor' inter-rater agreement. This is perhaps not an altogether surprising result considering that the questions were intentionally vague and

non-technical to allow for individual expert perspectives (the implications of which will be discussed in more detail in Chapter 4).

The individual expert scores for each of the 112 collocations were then totalled to give each collocation a total expert score on a standardised scale of 5-20 (the lowest total expert score a collocation could receive was 5 and the highest total expert score a collocation could receive was 20) (Table 16). The totalled expert scores were used as the basis of correlation analyses with the logDice scores (Table 16: correlation analysis 1) and the raw frequencies (Table 17: correlation analysis 2). The aim of these correlation analyses, which were carried out in Microsoft Excel, were to determine whether there was any correlation between pedagogical relevance (expert score) and statistical significance (logDice) and/ or pedagogical relevance (expert score) and frequency in academic texts (raw frequency). If there was strong correlation between, say, the expert scores and logDice scores, there might be evidence for an empirically-based weeding and ranking of the list.

Table 16. examples of individual and total expert (E) scores¹⁶ and logDice scores

Collocation	E1	E2	E3	E4	E5	Correlation analysis 1	
						Expert total	logDice
chapter-n aim-v	4	2	1	4	2	13	9.51
mutually-a exclusive-j	4	4	3	1	4	16	13.25
increasingly-a become-v	3	2	2	2	2	11	10.08
employ-v strategy-n	4	3	4	2	4	17	10.14
family-n home-n	3	2	2	1	2	10	10.02
social-j role-n	4	4	4	4	4	20	7.88

Table 17. Examples of individual and total expert (E) scores and raw frequency

Collocation	E1	E2	E3	E4	E5	Correlation analysis 2	
						Expert total	Raw freq.
chapter-n aim-v	4	2	1	4	2	13	28
mutually-a exclusive-j	4	4	3	1	4	16	41
increasingly-a become-v	3	2	2	2	2	11	38

¹⁶ For reasons of data protection and confidentiality, the expert numbers (e.g. E1, E2, etc.) in this table do not match the expert descriptions found earlier in this section.

employ-v strategy-n	4	3	4	2	4	17	45
family-n home-n	3	2	2	1	2	10	41
social-j role-n	4	4	4	4	4	20	38

The correlation analysis between the expert scores and logDice scores (correlation analysis 1) resulted in a Pearson Correlation Coefficient (PCC) (a coefficient which describes the association between two continuous variables) of -0.125. The correlation analysis between the expert scores and raw frequency (correlation analysis 2) resulted in a PCC of -0.138. According to Cohen's (1988) conventions for interpreting effect size in social science research, PCCs of less than -0.3 denote 'weak' negative correlation. Therefore, the present study found no meaningful correlation between either pedagogical relevance and statistical significance or pedagogical relevance and frequency in academic texts. These results do not provide any means by which to weed entries from the full list or rank remaining entries.

With the aforementioned correlation analyses not providing actionable results, the scores from the expert review were manually vetted to determine whether there existed a qualitative justification for the experts' judgements. This entailed rank ordering the 112 entries from the subset by their expert scores (from smallest to largest value) in Microsoft Excel and then scrutinising the data. This process of scrutinization revealed a pattern in relation to the 82 adjective-noun (J+N) combinations in the subset – the most common syntactic combination in the subset. There was a clear contrast between the adjectives in J+N combinations which were rated lowly by the experts and the adjectives in J+N combinations that were rated highly by the experts. For example, there is a clear difference between the adjectives *new*, *same* and *small* from the collocations *new way*, *same name* and *small group* (all with expert scores of 8) and the adjectives *cultural*, *dominant* and *literary* from the collocations *cultural capital*, *dominant narrative* and *literary genre* (all with perfect expert scores of 20) (Table 18).

Table 18. The lowest and highest rated adjective-noun combinations

Lowest rated J+N combinations		Highest rated J+N combinations	
Adjective + noun combination	Total expert score	Adjective + noun combination	Total expert score
new kind	8	literary tradition	19
new way	8	postmodern culture	19
same name	8	cultural capital	20
small group	8	dominant narrative	20
young man	8	literary genre	20

great deal	9	public sphere	20
large group	9	social order	20
large number	9	social role	20
modern man	9	social system	20
small number	9	visual representation	20

It seemed to the researcher that the adjectives in J+N combinations with lower expert scores (e.g. *same* and *small*) were more 'common' than the adjectives in combinations with higher expert scores (e.g. *liminal* and *postmodern*). That is to say, it appeared that what might distinguish the adjectives in J+N combinations with higher expert scores from those with lower expert scores was frequency in a general English corpus. To test this hypothesis, the raw frequencies of all the adjectives from the 82 J+N combinations in the subset were retrieved from the BNC. These raw frequencies were then used in a correlation analysis with the expert scores for the 82 J+N combinations (Table 19: Correlation analysis 3).

Table 19. Examples of expert scores and adjective frequency in BNC

Adjective + noun combination	Correlation analysis 3	
	Total expert score	Raw frequency of adjective in BNC
dominant -j culture-n	18	3,001
dominant -j narrative-n	20	3,001
golden -j age-n	17	2,777
raw -j material-n	14	2,355
following -j decade-n	13	1,952
explicit -j reference-n	17	1,867
theatrical -j performance-n	17	592
facial -j expression-n	12	536
postmodern -n culture-n	19	151
liminal -j space-n	18	16

The resulting PCC was -0.575, which, according to Cohen's (1988) conventions, denotes strong negative correlation. In other words, the more 'common' the adjective in a J+N combination, the less pedagogically relevant the experts deemed the combination to be. It was, therefore, decided that J+N combinations with a 'common' adjective would be removed from the full list. This had already been attempted (unsuccessfully) during the noise removal

process, though at this point it was felt that it could be done empirically. Accordingly, the 82 J+N combinations from the subset were ordered by the frequency of their adjective in the BNC from highest to lowest (Table 20) and closely examined in order to find a suitable value threshold at which adjectives could be considered ‘common’ and weeded from the list.

Table 20. Top 28 (of 82) Adjective-noun combinations ordered by the frequency of the adjective in the BNC (from highest to lowest)¹⁷

Adjective + noun combination	Totalled expert score (out of 20)	Raw freq. of adjective in BNC (highest to lowest)
new-j way-n	8	105,682 (most common)
new-j model-n	14	105,682
new-j kind-n	8	105,682
new-j genre-n	17	105,682
same-j name-n	8	61,160
long-j time-n	10	60,787
great-j length-n	13	59,694
great-j deal-n	9	59,694
great-j detail-n	13	59,694
old-j generation-n	11	58,607
small-j number-n	9	50,816
small-j group-n	8	50,816
different-j kind-n	11	47,546
different-j way-n	11	47,546
different-j form-n	13	47,546
large-j number-n	9	47,415
large-j group-n	9	47,415
local-j community-n	14	45,361
early-j day-n	13	40,920
early-j stage-n	13	40,920
present-j time-n	13	36,515
social-j class-n	16	36,511
social-j system-n	20	36,511

¹⁷ Table 20 does not represent the full data set – there are 82 J+N combinations and this represents only the top 28 ordered by adjective frequency in the BNC.

social-j role-n	20	36,511
social-j order-n	20	36,511
public-j sphere-n	20	36,144
further-j development-n	16	35,913
young-j man-n	8	35,458 (less common)

Finding a suitable value threshold for ‘commonness’, though, proved extremely difficult and it became apparent that it would be problematic for two key reasons. Firstly, some J+N combinations containing the same common adjective received vastly different expert scores. For example, J+N combinations containing the common adjective *new* (BNC frequency of 105,682), received expert scores ranging from 8 (*new way*) to 17 (*new genre*) (Table 20). Secondly, some J+N combinations with low expert scores contained less common adjectives than J+N combinations with high expert scores. For example, the adjective *young* (BNC frequency of 35,458) is less common than the adjective *social* (BNC frequency of 36,511), yet *young man* received an expert score of 8 while *social system*, *social role* and *social order* all received perfect expert scores of 20 (Table 20). For these reasons, it was clear that any weeding of ‘common’ adjectives would also remove entries of pedagogical relevance. Therefore, despite evidence to suggest that J+N combinations containing common adjectives tend to receive low expert scores, there was no reliable method by which to remove these common adjectives because in certain collocational relationships common adjectives take on more technical meanings (e.g. *new order* and *public sphere*).

In sum, then, there was a ‘poor’ level of inter-rater agreement between the five experts (ICC 0.315) (which will be discussed in Chapter 4). Furthermore, the associations between expert scores and logDice scores (PCC -0.125), expert scores and raw frequency (PCC -0.138) and expert scores and adjective ‘commonness’ (PCC -0.575) are not strong enough or reliable enough to enable an empirical weeding or ranking of the full list (which will also be discussed in Chapter 4). Therefore, no entries were removed by the expert review process and the final ACLAH contains 555 collocations.

3.9. Presenting the ACLAH

In order to present the final ACLAH in a systematic and user-friendly manner, the final 555 collocations, listed as lemmas plus POS suffixes, were categorised by their syntactic combination and ordered alphabetically. As discussed above, it was not possible to order the entries by pedagogical relevance, yet alphabetical ordering does facilitate teaching-learning with the ACLAH. It allows recurrent collocational frameworks to be easily identified and

targeted, e.g. *offer an example*, *offer an insight* and *offer a perspective* – these entries might be separated if they were ranked by pedagogical relevance. To further facilitate future classroom (and research) use, alongside each ACLAH entry are range, frequency and logDice statistics and the ‘longest-commonest match’ (Kilgarriff et al., 2013) (Table 21). The longest commonest match is particularly useful in that it reveals the most common realisation of a collocation – an important pedagogical consideration when working with lemmas rather than words (see 3.2.1) (see Appendix 2 for full ACLAH).

Table 21. Presentation of the final ACLAH

(SUBJ.) N + V Combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCOMP	FILMTV	MODLANG	THEATRE			
actor-n play-v	14	12	7	34	67	10.3	<i>actor playing</i>
audience-n see-v	12	12	2	14	40	8.98	<i>audience sees</i>
chapter-n aim-v	6	6	6	10	28	9.51	<i>this chapter aims to</i>

4. Results and discussion

In this chapter, the ACLAH will be evaluated. The similarities and differences between the composition of the ACLAH and the ACL will be discussed (4.1), and then a validation study will be carried out in order to assess the usefulness of the ACLAH, particularly in comparison with the ACL (4.2). The aim of these two subsections is to address the following questions:

- (1) How is the ACLAH similar and/ or different to the ACL?
- (2) Is the ACLAH equally useful across Arts and Humanities fields?
- (3) Is the ACLAH a readily usable resource for EAP teaching-learning purposes?

4.1. Composition of the ACLAH

In terms of syntactic combinations, the overall composition of the ACLAH is extremely similar to that of the ACL. What is particularly interesting is that in both lists the most common syntactic relationship is by far adjective-noun combinations (69% of the ACLAH and 71.8% of the ACL) (Table 22). What is more, in the present study, the only items to receive perfect expert scores were adjective-noun combinations (Table 23), which demonstrates a high-level of agreement amongst EAP practitioners that certain adjective-noun combinations are highly pedagogically relevant. However, this syntactic combination, which clearly dominates academic register, has not hitherto attracted much research interest, particularly in comparison with verb-noun combinations. As Ackerman and Chen (2013:240) state, ‘the assumption appears to be that learners tend to encounter more difficulties in choosing verb collocates correctly than any other type of collocation’. Yet, in one of the few studies regarding adjective-noun combinations, Siyanova and Schmitt (2008) reported that 25% of adjective-noun combinations produced by learners were atypical (e.g. *plastic operation* rather than *plastic surgery*). Adjective-noun combinations, then, perhaps require more attention in EAP research and EAP teaching-learning.

Table 22. Syntactic combinations of the ACLAH

Syntactic combination	No. of entries	Percentage of the ACLAH	Examples
Adjective + noun (J+N)	383	69%	active participation, basic principle, civil society
Noun + noun (N+N)	68	12.3%	research questions, source material, time period
Subject noun + verb (N+V)	8	1.4%	chapter focus, film show, audience see
Verb + object noun (V+N)	41	7.4%	blur line, challenge notion, follow model

Adverb + adjective (A+J)	12	2.2%	mutually exclusive, slightly different, highly influential
Verb + adverb (V+A)	43	7.7%	already mention, clearly define, inextricably link

Table 23. Collocations with the highest expert scores

Collocations	E1	E2	E3	E4	E5	Total
dominant-j narrative-j	4	4	4	4	4	20
social-j role-n	4	4	4	4	4	20
social-j order-n	4	4	4	4	4	20
cultural-j capital-n	4	4	4	4	4	20
visual-j representation-n	4	4	4	4	4	20
social-j system-n	4	4	4	4	4	20
literary-j genre-n	4	4	4	4	4	20
public-j sphere-n	4	4	4	4	4	20

In contrast with the striking similarity discussed above, there is a striking difference between the composition of the ACLAH and the ACL. While almost half of ACLAH entries (264 out of 555) are combinations of two items from the GSL (West, 1953) (e.g. *critical discussion* and *directly address*), ACL entries are constrained to only one GSL item per entry (e.g. *alternative model* and *assume responsibility*). This is because, in an attempt to identify ‘academic’ collocations, Ackerman and Chen (2013:237) ostensibly excluded GSL items from occurring as node words¹⁸. Yet, it is clear from the present study that certain GSL items can take on technical meanings in collocational relationships (Fig. 11), and these relationships are prevalent in AH discourse. If these relationships are as prevalent in general academic discourse, the ACL potentially overlooks much that might be of collocational interest to EAP by disallowing the co-occurrence of two GSL items. Therefore, it seems that any future attempts to compile lists of academic collocations would be better served by adopting the present study’s approach to identifying ‘academic’ collocations through keyness (as discussed in 3.2.4).

Fig. 11. ACLAH collocations containing two GSL items

¹⁸ Ackerman and Chen (2013:237) clearly state that ‘words from the General Service List were also removed from the node words list but could appear as pre- or post-collocate’. However, a close inspection of the ACL using the online GSL highlighter reveals that there are in fact instances of two GSL words forming an entry in the ACL. These instances are infrequent (out of the first 250 entries in the ACL only 22 are combinations of two GSL words), yet nowhere in their paper do Ackerman and Chen acknowledge or explain how these entries came to be on the ACL.

new order	bring together	high art	draw together
body (of) work	everyday practice	popular audience	raw material
social value	shoot (a) film	broad sense	point (of) reference
political right	give voice (to)	critical framework	power relation
social practice	give sense (to)	employ (a) term	golden age

The ACLAH, then, represents progress towards a more comprehensive account of collocation in academic prose because it allows entries to be combinations of two GSL items, which are particularly prevalent in AH discourse. Yet, the ACLAH is not entirely problem free - it is composed of a number of entries which have been deemed pedagogically irrelevant by EAP practitioners. This is because the expert review did not yield strong enough agreement or correlation to provide a reliable method by which to weed the list. To have provided actionable results, the expert review would need to be improved in two ways: focused technical questions; and review of all 555 collocations.

Firstly, the questions asked in the expert review were intentionally vague and non-technical to allow for the incorporation of individual perspectives and different approaches to collocation (Fig. 12), which, it seems, resulted in poor agreement among the experts. In contrast, focused technical questions may have facilitated a higher level of agreement. For example, the experts could have been asked to classify collocations by their degree of fixedness with the intention of including only restricted collocations in the final listing (Fig. 13). This would require the experts to have *extensive* background knowledge of collocation literature, which *may* have improved the level of inter-rater agreement, but would ultimately have made the expert review time consuming beyond the constraints of the study.

Fig. 12. Vague expert review questions

Score each entry in consideration of the following two questions:

(a) Is it appropriate to consider the entry as an academic collocation for ESAP teaching-learning purposes?

(b) Do you think the collocation is worth teaching explicitly as part of an Arts and Humanities ESAP course?

Fig. 13. Focused technical expert review questions

Score each entry in consideration of the following question:

(a) Is one component used in a figurative, delexical or technical sense?

(b) Is commutability arbitrarily restricted?

Secondly, a full review of all 555 entries would have been beneficial because, as previously discussed, correlation analysis does not take sufficient account of the nuances of language and therefore cannot be reliably used to filter a list of collocations. For example, the strong

negative correlation between the commonness of an adjective and pedagogical relevance (PCC -0.575) failed to take into account the differences between collocations such as *new kind* and *new order* or *high level* and *high art*, which would have been treated as equally irrelevant and dismissed due to their shared ‘common’ adjectives. For this reason, future attempts to compile lists of academic collocations should adopt the approach of Ackerman and Chen (2013) and have all entries reviewed by the expert panel to allow for decisions to be made based on the nuances of individual entries. Again, though, this would have been beyond the temporal constraints of the present study.

It is, however, important to note that the expert review of 112 collocations, imperfect though it was, provided evidence to suggest that if all 555 entries had been reviewed by the experts, the overall composition of the final list would not have been drastically affected. Out of the 112 entries in the subset, only 11 (10%) received a total expert score of less than or equal to 9 (Table 24). Ackerman and Chen (2013:240), whose expert scores also operated on a standardised scale of 5—20¹⁹, excluded all entries from the ACL with a total expert score of less than or equal to 9. Using the same value threshold in the present study suggests that only approximately 55 entries (10%) would be removed if all 555 items were reviewed by the experts. This figure should be considered merely indicative, yet it is a promising result. It implies that the stages of computational analysis and manual refinement have yielded a list of collocations of which up to 90% are pedagogically relevant. The ACLAH, then, if used pragmatically, can be considered a readily usable resource for EAP teaching-learning (which will be discussed in more detail in Chapter 5).

Table 24. Collocations with the lowest expert scores (equal to or less than 9)

Collocations	E1	E2	E3	E4	E5	Total
experience-n (of) time-n	1	1	4	2	1	9
small-j number-n	4	1	1	1	2	9
large-j group-n	4	2	1	1	1	9
modern-j man-n	3	1	2	2	1	9
great-j deal-n	2	1	1	1	4	9
large-j number-n	3	2	1	2	1	9
young-j man-n	3	1	2	1	1	8

¹⁹ Ackerman and Chen’s (2013) panel included the same number of experts (5) using the same four-point Likert scale as present study (n.b. there were 6 experts in their panel but the ratings of one expert had to be disregarded as they were incomplete (ibid:240).

new-j kind-n	3	2	1	1	1	8
same-j name-n	1	1	3	1	2	8
small-j group-n	3	2	1	1	1	8
new-j way-n	3	2	1	1	1	8

4.2. Validation

A validation study was conducted to investigate the ACLAH's coverage of various corpora. The first validation analysis was carried out using the ACLAH's source corpus – the AHC – to determine whether the ACLAH was equally useful across the four fields of AH for which it was created. Calculating the ACLAH's coverage of the AHC was very straightforward because the computational analysis and manual exploration of dispersion yielded frequency data for all ACLAH entries which could be used to calculate coverage. For this calculation, ACLAH entries are classed as pairs of lemmas (including their inflected forms) co-occurring within a span of +/-5 (Fig. 14).

Fig. 14. ACLAH entries

ACLAH entries are classed as pairs of lemmas co-occurring within a span of +/- 5. For example, the ACLAH entry *focus-v specifically-a* encompasses the following positionally and inflectionally variable occurrences:

Node *focus* with collocate *specifically* in span position -1

specifically-a focus-v
specifically-a focuses-v
specifically-a focusing-v
specifically-a focused-v

Node *focus* with collocate *specifically* in span position +1

focus-v specifically-a
focuses-v specifically-a
focusing-v specifically-a
focused-v specifically-a

Node *focus* with collocate *specifically* in span position +2

focus-v (more) specifically-a
focuses-v (more) specifically-a
focusing-v (more) specifically-a
focused-v (more) specifically-a

The 555 entries ACLAH entries occur in the AHC as pairs of lemmas within a span of +/- 5 a total of 30,256 times, which amounts to a total coverage of 0.54% (Table 25). This figure is as

expected considering that the ACL, with four times as many entries, covers 1.4% of its source corpus (Ackerman and Chen, 2013:243)²⁰. What is more important here, though, is the coverage of the subcorpora which each represent a different field within AH. The ACLAH covers between 0.41% and 0.76% of the four subcorpora (Table 25). This margin may seem negligible as a percentage, but when the average normed frequency of each entry is calculated, it becomes clear that it is quite a significant difference. For example, in the ENGCOMP subcorpus each ACLAH entry occurs on average 7.3 times p/m words, whereas in the THEATRE subcorpus each entry occurs on average 13.7 times p/m words – almost double. This result has implications for teaching and learning with the ACLAH (which will be discussed in Chapter 5).

Table 25. ACLAH coverage of AHC and its subcorpora

AHC and subcorpora	Word count	ACLAH occurrences	Coverage percentage
ENGCOMP	~1,509,793	6,230	0.41%
FILMTV	~1,526,945	8,725	0.57%
MODLANG	~1,372, 451	6,354	0.46%
THEATRE	~1,178,695	8,947	0.76%
AHC (TOTAL)	5,587,887	30,256	0.54%

The second validation analysis was carried out to determine whether students in AH would be better served by the ACLAH or the ACL. In order to do this, the ACL's coverage of the AHC would need to be calculated and compared to the ACLAH's coverage of the AHC. The usual way to calculate the coverage of a corpus is to use a Whitelist - a list of vocabulary in the format of a .txt file which, when selected in the Sketch Engine's Word List tool, enables the frequency data for all items on the list to be gathered in one search. A Whitelist search, though, cannot gather frequency data for *collocations* (words in a relationship that may or may not be next to each other), but it can gather frequency data *for n-grams* (words that are directly next to each other). It was using n-grams that Ackerman and Chen (2013:241) calculated the ACL's coverage of its source corpus, and therefore how the ACL's coverage of the AHC will be calculated. This means that, *solely* for the purposes of this analysis, entries are classed as pairs of lemmas (including their inflected forms) co-occurring directly next to each other (Fig. 15).

²⁰ It is important to note that that, unlike the ACLAH, the ACL's validation analysis of coverage did not include positional variability and therefore the percentage could in fact be slightly higher.

Fig. 15. ACLAH entries as n-grams

N-grams are pairs of lemmas co-occurring directly next to each other. For example, as an n-gram, the ACLAH entry *focus-v specifically-a* encompasses only the following occurrences:

Node *focus* with collocate *specifically* in span position +1

focus specifically

focuses specifically

focusing specifically

focused specifically

To guarantee a fair comparison, both the ACLAH's coverage and the ACL's coverage of the AHC would need to be calculated as n-grams (because the ACLAH's total coverage figure of 0.54% includes positionally variable entries within a span of +/-5). Therefore, n-gram lists for both the ACLAH and the ACL were created in which all entries were listed as lemma²¹ without any articles, copula 'be' or prepositions²². These lists were then used as Whitelists in the Sketch Engine to calculate the two lists' coverage of the AHC (Screenshot 9).

Screenshot 9. Using a Whitelist to generate frequency data for an n-gram list

²¹ This required that VPP (verb past participle) entries on the ACL were changed to the verb in its base form, e.g. *strongly-a influenced-vpp* was changed to *strongly-a influence-v* (which encompasses *strongly influenced*).

²² ACL entries, such as *explore (an) issue*, *(be) particularly evident* and *inextricably linked (to, with)* were changed to *explore issue*, *particularly evident* and *inextricably link* to conform with ACLAH entries and ensure that items such as *an*, *be*, *to* and *with* were not included in the n-gram analysis.

Word list options

Subcorpus: None (whole corpus) [info](#) [create new](#)

Search attribute: lemma (lowercase)

use n-grams. Value of n: from 2 to 2

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression:

Minimum frequency: 1

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: ACL_ngrams.txt

Blacklist: No file chosen [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple

Keywords

Reference (sub)corpus: English Web 2013 (enTenTen13) (whole corpus)

Prefer: rare words common words 1

Change output attribute(s)

--- --- ---

You can select one or more output attributes. Please note that this option can be time-consuming.

The results of this analysis are striking. The 2,468 n-grams of the ACL occur 19,705 times in the AHC, equating to an overall coverage of 0.35%, whereas the 555 n-grams of the ACLAH occur 25,906 times, equating to an overall coverage of 0.46%²³ (Table 26). To put this into perspective, each ACL n-gram occurs on average 1.4 times p/m words, whereas each entry ACLAH n-gram occurs on average 8.4 times per million words. So, not only does the ACLAH provide a higher coverage *but* each entry occurs on average 6 times more frequently than its ACL counterpart. This clearly demonstrates that the collocational needs of EAP students in the four fields of AH comprising the AHC would be much better served by the ACLAH than the ACL.

Table 26. ACLAH and ACL n-gram coverage of the AHC

List	AHC (5,587,887 total words)		
	Total occurrence for all items	Average occurrence per entry	Total coverage
ACL	19,705	8	0.35%
ACLAH	25,906	47	0.46%

²³ It should be noted that the ACLAH entries as *collocations* cover 0.54% while the ACLAH entries as *n-grams* cover 0.46%. The difference of only 0.08% suggests that n-gram coverage is a fairly good indicator of collocation coverage (likely due to the high number of adjective-noun combinations which do not exhibit much positional variability).

It could, though, be argued that the above analysis is unfairly weighted in favour of the ACLAH. As Coxhead (2000:224) states ‘a frequency-based word list that is derived from a particular corpus should be expected to cover that corpus well [and therefore] the real test is how the list covers a different collection of similar texts’. To that end, a third validation study was carried out using the BAWE’s preloaded 1.9 million-word AH subcorpus to represent ‘a different collection of similar texts’²⁴. The two n-gram lists were used to calculate the ACLAH’s and ACL’s coverage of the BAWE AH subcorpus.

The ACL’s coverage of the BAWE AH is 0.39%, whereas the ACLAH’s is 0.3% (Table 27). Yet, it is important to observe that due to the vast size difference between the lists, each ACL entry occurs on average only 1.6 times p/m words, while each ACLAH entry occurs 5.5 times p/m million words. This is particularly interesting because, although the ACLAH was created for a very specific subdivision of AH, it could be argued that, despite the overall coverage difference of 0.09%, the ACLAH, with fewer entries and more occurrences p/m words, is more ecologically valid for EAP teaching-learning in ‘general’ AH than the ACL.

Table 27. ACLAH and ACL n-gram coverage of the BAWE AH

List	BAWE AH (1,875,147 total words)		
	Total occurrence for all items	Average occurrence per entry	Total coverage
ACL	7,226	3	0.39%
ACLAH	5,699	10	0.30%

²⁴ The BAWE Arts and Humanities (AH) subcorpus subsumes Linguistics, Archaeology Classics, Comparative American Studies, History and Philosophy. This is in some ways similar and in other ways dissimilar to the AHC because, as previously discussed, there are many ways to define AH.

5. Conclusions and implications

The ACLAH is a list of 555 pedagogically useful academic collocations intended for use as a teaching-learning resource in EAP, moreover ESAP. It is the result of a mixed-method approach to collocation analysis which combines computational corpus linguistics typical of the neo-Firthian approach with manual intervention typical of the phraseological approach. Computational analysis provided the means by which to quantitatively identify collocations which are significantly more frequent in AH texts than general English texts, while manual intervention provided the means by which to qualitatively filter the collocations to ensure the final list provides a readily usable EAP resource.

The ACLAH is different to existing listings in that it does not overlap with what is traditionally defined as ‘academic’ vocabulary. Almost half of ACLAH entries are the combination of two GSL items and only 130 ACLAH entries are common to the ACL (Fig. 16). This is largely due to the present study’s methodological approach to identifying ‘academic’ vocabulary through keyness, moreover Simple Math (Kilgarriff, 2009), which allows GSL items to occur without restriction in the ACLAH. As discussed and demonstrated in this study, high-frequency items in collocational relationships are ubiquitous in AH discourse and not always as semantically transparent as their individual counterparts might suggest. Although budding EAP learners will likely be familiar with individual GSL items like *employ* and *shoot*, combinations such as *employ (a) term* and *shoot (a) film* ‘probably require specific pedagogical attention’ (Durrant, 2009:164). Therefore, by allowing for these combinations, which are mostly overlooked by the ACL, the ACLAH represents progress towards a more comprehensive account of academic collocation.

Fig. 16. Collocations common to both the ACLAH and the ACL

active participation	critical analysis	crucial role	detailed analysis	private sphere
common culture	certain aspect	local community	offer insight	basic principle
cultural life	historical account	traditional value	privileged position	civil society
cultural background	original meaning	historical period	key element	integral part
new perspective	literary text	central concern	historical context	popular culture
political context	modern society	key factor	religious belief	highly influential
cultural context	traditional view	final chapter	close relationship	mental state
cultural history	capitalist society	explore issue	previous section	digital technology
cultural practice	key role	complex relationship	cultural diversity	public sphere
historical change	social context	visual representation	historical event	facial expression
physical space	literary tradition	cultural identity	sexual difference	slightly different
traditional form	collective identity	thought process	legal system	national identity
political reality	significant role	contemporary society	young generation	ethnic group
make argument	key issue	central role	defining feature	vast majority

traditional culture	key aspect	prime example	particularly relevant	textual analysis
previous work	social status	physical appearance	emotional response	source material
cultural heritage	crucial point	critical attention	creative process	well aware
cultural tradition	further development	high level	further evidence	previous chapter
social background	human activity	dominant culture	particularly evident	binary opposition
national culture	natural world	central theme	public discourse	culturally specific
visual image	specific context	previous decade	collective memory	stark contrast
social structure	personal experience	primary source	fully understand	mutually exclusive
critical approach	essential element	special issue	support argument	national boundary
use strategy	original text	critical perspective	domestic sphere	final section
well received	directly linked	clearly defined	previously discussed	dominant ideology
primarily concerned	closely related	closely associated	inextricably linked	

The most significant difference between the ACLAH and existing vocabulary lists, though, is that rather than misrepresenting academic literacy as a uniform practice, the ACLAH engages with current conceptions of academic literacies by acknowledging that different collocations occur and behave differently across different disciplines (Hyland and Tse, 2007). As discussed by Durrant (2009) and demonstrated in this paper, the vocabulary needs of AH students are very different from those in other disciplines. For example, in the ACLAH's source corpus (the AHC), each ACLAH entry occurs on average 8.4 times p/m words, while each ACL entry only occurs on average 1.4 times p/m words²⁵. What is more, even in the BAWE AH subcorpus, a different collection of similar texts, each ACLAH entry occurs on average 5.5 times p/m words, while each ACL entry only occurs on average 1.6 times p/m words²⁶. These findings add to the mounting evidence casting doubt on the notion of a 'core' academic vocabulary. Moreover, they highlight the usefulness of the ACLAH and reinforce the need for more specific listings of academic collocations to be compiled for ESAP teaching-learning purposes.

In terms of EAP teaching-learning, it is hoped that the ACLAH will be integrated into a lexical syllabus or used as the basis of a lexical unit within a traditional syllabus. The ACLAH can be used to set vocabulary goals, design teaching materials and draw students attention to useful collocations (Coxhead, 2000:228). The explicit teaching of the ACLAH is facilitated by its presentation which is subdivided by syntactic combination and includes frequency data. The subdivisions can be used to set manageable vocabulary learning goals during a course of study. For example, teachers and students can easily direct more attention to, say, adjective-noun combinations which, as highlighted, dominate academic register. The frequency data

²⁵ This figure is not absolute as it is based on n-gram coverage rather than collocation coverage.

²⁶ As above

enables teachers and students to make informed decisions about which collocations might be most useful in their particular fields of AH. This is especially important because, as revealed in the validation study, items behave differently in terms of frequency across fields of AH. Particularly in the field of ENGCOMP, where ACLAH entries occur less frequently than in other fields, teachers and students might want to use the frequency data to prioritise entries.

Explicit teaching, though, needs to be mixed with opportunities for implicit learning. Students must be afforded opportunities to meet the collocations in message-focused reading and listening and to use the collocations in speaking and writing (Coxhead, 2000:228). Although the ‘longest-commonest match’ provides some context which may significantly aid learners in understanding the most typical use of each collocation, it is not sufficiently message-focused. Rather, concordancing software, such as the Sketch Engine, provides teachers the means by which to identify ACLAH items in message-focused texts. For example, using a Whitelist in the Sketch Engine’s Word List tool, all 383 ACLAH adjective-noun combinations can reliably be identified as n-grams in a text²⁷ and subsequently highlighted or even removed as the basis of a gap fill. Furthermore, students themselves can be encouraged to use concordancing software to complete data driven learning activities. For example, students can be encouraged to use concordance lines to induce the specialised meanings of ACLAH entries such as *golden age* (Screenshot 10). In sum, a blend of explicit and implicit learning may significantly contribute to the acquisition of this discipline-specific set of vocabulary (Wang, Liang and Ge, 2008).

Screenshot 10. Concordance lines for ‘golden age’

²⁷ The n-gram method can reliably identify adjective-noun combinations because they exhibit little positional variability

Query 42 (6.15 per million) ⓘ

1 ENGCMP_PH... 358), and along with its fabled golden city, the **golden age** is poignantly and irrevocably gone. It is no
2 ENGCMP_PH... accessible fruits of what has been termed the "**golden age**" for the study of medieval literary theory
3 ENGCMP_PH... xxiv), but its upper limit does not extend to the **golden age** of Elizabethan drama: "the Renaissance
4 ENGCMP_PH... life in values which were most important in the **golden age** of bourgeois society, like family values,
5 ENGCMP_PH... poetry, written mostly in MiddleChinese in the **golden ages** of Han (206 BC-AD 220), Tang (AD 618-907)
6 FILMTV_PHD... Mann and the digitalOpening in 1933-"the **golden age** of bank robbery," as the opening titles
7 FILMTV_PHD... , Alvin Karpis and Baby Face Nelson itis the **golden age** of bank robbery..."191 often demonstrated
8 FILMTV_PHD... from 1930 to mid-1934 as 'juricinema's first **golden age**',6 identifying a distinct pattern of trial
9 FILMTV_PHD... the films of American trial cinema's so-called **golden age** of 1957-62: Twelve Angry Men (Dir: Sidney
10 FILMTV_PHD... cinema in America',36 and the legal film's '**golden age**'.37 Papke interrogates the shared
11 FILMTV_PHD... the shared ideology underlying these "**golden age**" images of U.S. law, although he notes the
12 FILMTV_PHD... form. Papke refigures Nevins' notion of a "**golden age**" to point to the films' shared ideological
13 FILMTV_PHD... his subsequent attempt to historicise this **golden age** as part of Hollywood's assertion of its
14 FILMTV_PHD... of social and cultural factors influenced the **golden age** of the trial film in the late-1950s. 49 Ibid.
15 FILMTV_PHD... I would posit understandably) absent from the "**golden age**" corpus, warrants further explanation
16 FILMTV_PHD... constitute what has been figured as the **golden age** of the trial film. These films include
17 FILMTV_PHD... and differences identifiable across the **golden age** legal films.103 All but two of the films take
18 FILMTV_PHD... appears more frequently in the American-set **golden age** criminal trial films. The British court of
19 FILMTV_PHD... , p. 59. 71 'flat' depiction. Papke argues that **golden age** 'characterisation depends on a political "
20 FILMTV_PHD... the heroic defence roles in the majority of the **golden age** legal films, prosecution lawyers are
21 FILMTV_PHD... tone whilst departing drastically from the **golden age** representational strategies. This
22 FILMTV_PHD... of trial material to television. Although the **golden age** of 1957-62 seemed to incorporate both film
23 FILMTV_PHD... the past, with the result that the idealised **golden age** represented innostalgia texts could
24 FILMTV_PHD... idealise the past, aseither a lost "**golden age**" or in glossy pastiche that covers over the
25 FILMTV_PHD... period after McCarthy and beforeVietnam as a '**golden age**' of youth culture.809 Although set only
26 FILMTV_PHD... that invites its interpretation as an idyllic "**golden age** ." Pointingto the difficulty in regarding
27 FILMTV_PHD... Japanese directors. The period 1969-1981 was a **golden age** of Suzuki criticism, with several feature
28 MODLANG_PH... , endowing it with traits distinctive of the **golden age** . Up to the very last lines, the reader is
29 MODLANG_PH... , a period that has been referred to as the '**golden age** of outre-mer travel accounts'.70 Accounts
30 MODLANG_PH... changes in an attempt to save the remnants of a **golden age** , 55Henri Rossi, Mémoires aristocratiques
31 MODLANG_PH... . The tension between the wish to testify of a **golden age** overcomes the fear of becoming a female
32 MODLANG_PH... , hinting at Boigne's despair to preserve a **golden age** . Boigne was not the only one to worry about

Finally, it is important to point out a caveat of the ACLAH. Because the expert review did not provide actionable results, a number of pedagogically questionable entries remain on the list. Consequently, although there is evidence to suggest that up to 90% of ACLAH entries are pedagogically relevant, users must be pragmatic in their approach to individual entries. That is to say, as with any vocabulary list (particularly one with, say, 2468 entries), teachers will need to make principled decisions about what ACLAH content to draw students' attention to for maximum benefit. A revised ACLAH would most definitely seek to have all 555 entries reviewed by a panel of experts in order to reliably remove the small number of accretions such as *young man* and *same name*.

References

- Aisenstadt, E.** (1981). *Restricted Collocations in English Lexicology and Lexicography*. *ITL Review of Applied Linguistics* 53: 53-61.
- Aitchison, J.** (1994). *Words in the mind: an introduction to the mental lexicon* (2nd ed). Blackwell, Oxford, UK.
- Bahns, J.** (1993). *Lexical collocations: A contrastive view*. *ELT Journal*, 47(1), 56–63.
- Bahns, J., & Eldaw, M.** (1993). *Should we teach EFL students collocations?* *System*, 21(1), 101–114.
- Benson, M.** (1985). *Collocation and idioms*. In R. Ilson (Ed.), *Dictionaries, lexicography and language learning* (pp. 61–68). Oxford: Pergamon Press.
- Benson, M., Benson, E. & Ilson, R.** (1997) *The BBI dictionary of English word combinations*. Amsterdam, The Netherlands: John Benjamins Publishing Company
- Becher, T.** (1989) *Academic Tribes and Territories: Intellectual enquiry and the culture of disciplines*. Buckingham: Open University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Biber, D., Conrad, S., & Cortes, V.** (2004). If you look at.: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.
- Boers, F., Eyckmans, J. Kappel, H. Stengers, & Demecheleer, M.** (2006). *Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test*. *Language Teaching Research*, 10, 245–261.
- Bolinger, D.** (1976). *Meaning and memory*. *Forum Linguisticum* I: 1-14.
- Campion, M. E., & Elley, W. B.** (1971). *An academic vocabulary list*. Wellington, New Zealand: New Zealand Council for Educational Research.
- Chung, T. & Nation, I. S. P.** (2003). *Technical Vocabulary in Specialised Texts*. *Reading in a Foreign Language*. 15:2, 103-116

- Cicchetti, D.V.** (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, **6**(4): 284–290.
- Clear, J.** (1993). From Firth principles: Computational tools for the study of collocations. In **M. Baker, G. Francis, & E. Tognini-Bonelli** (Eds.). *Text and technology: In honour of John Sinclair* (pp. 271–292). Amsterdam: John Benjamins.
- Cohen, J.** (1988). *Statistical Power Analysis for the Behavioral Sciences*. USA: Routledge.
- Cobb, T.** (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, **59**(3), 393–423.
- Cowie, A. P.** (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, **2**(3), 223–235.
- Cowie, A. P.** (1998). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press.
- Coxhead, A.** (2000). A new academic word list. *TESOL Quarterly*, **34**(2), 213–238.
- Coxhead, A.** (2008). Phraseology and English for academic purposes. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149–161). Amsterdam: John Benjamins.
- Durrant, P.** (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, **28**(3), 157–169.
- Eldridge, J.** (2008), "No, There Isn't an 'Academic Vocabulary,' But...": *TESOL Quarterly*, **42**: 109-113.
- Ellis, N.** (2003). Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. In **C. J. Doughty, & M. Long** (Eds.), *The Handbook of Second Language Acquisition* (pp. 63-103). Malden, MA: Blackwell.
- Ellis, N. C., Simpson-Vlach, R., & Maynard, C.** (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, **41**(3), 375–396.
- Firth, J. R.** (1957). *Papers in linguistics 1934–1951*. London: Oxford University Press.

- Foucault, M.** (1981). *The order of discourse*. In **R. Young** (Ed) (1981), *Untying the text: a post-structural anthology* (pp. 48- 78). Boston: Routledge & Kegan Paul.
- Gablasova, D., Brezina, V. & McEnery.** (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(Suppl. 1), 130-154.
- Gardner, D. and Davies, M.** (2014). A new academic vocabulary list. *Applied linguistics*, 35(3), 305-327.
- Gledhill, C.** (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.
- Hausmann F. J.** (1989). *Le dictionnaire de collocations*. In **Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L.** (eds). *Wörterbücher: ein internationales Handbuch zur Lexicographie*. Berlin/New-York, De Gruyter, 1010-1019.
- Hermann, M.** (2015). *Cluster Analysis for Corpus Linguistics*. Berlin; Boston: De Gruyter.
- Hoey, M.** (1991). *Patterns of lexis in Text*. Oxford: Oxford University Press.
- Hoey, M.** (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
- Hoey, M.** (2007). *Lexical priming and literary creativity*. In **Hoey, M., Mahlberg, M., Sinclair, J., Stubbs, M. & Teubert, W.** (2007). *Text, Discourse and Corpora*. London; New York: Continuum.
- Howarth, P.** (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen: Niemeyer.
- Hunston, S.** (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Hyland, K., & Tse, P.** (2007). Is there an “Academic vocabulary”? *TESOL Quarterly*, 41(2), 235–253.
- Hyland, K. (2008).** *As can be seen: Lexical bundles and disciplinary variation*. *English for Specific Purposes*. 27, 4-21.
- Hyland, K.** (2012). *Disciplinary Identities: Individuality and Community in Academic Discourse*. Oxford University Press.

Kilgarriff, A. (2009). *Simple maths for keywords*. In *Proceedings of Corpus Linguistics Conference CL2009*, Mahlberg, M., González-Díaz, V. & Smith, C. (eds.), University of Liverpool, UK, July 2009.

Kilgarriff, A., Baisa, V., Rychly, P. & Jacubicek, M. (2018) *Longest-commonest match*. In **I. Kosem, M. Jakubiček, J. Kallas, and S. Krek** (eds.): *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom: 397–404. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Clarendon Press.

Laufer, B. (1991). *How much lexis is necessary for reading comprehension?* In **P. Arnaud, & H. Bejoint** (Eds.), *Vocabulary and applied linguistics* (pp. 316–323). London: Macmillan.

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography*, 24(1), 29–49.

Laufer, B., & Sim, D. D. (1985). Measuring and explaining the threshold needed for English for academic purposes texts. *Foreign Language Annals*, 18, 405–413.

Leech, G. (1974). *Semantics*. Harmondsworth: Penguin.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications

Lewis, M. (Ed.) (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: LTP.

Li, Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System*, 38, 402-411.

Louw, B. (1993). *Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies*. In **M. Backer, G. Francis, & E. Tognini-Bonelli** (Eds.), *Text and technology* (pp. 157–176). Amsterdam: Benjamins.

Lyons, J. (1977). *Semantics: Vol. 1*. Cambridge University Press.

Martinez, I. A., Beck, S.C. & Panza, C.B. (2009). *Academic vocabulary in agriculture research articles: A corpus-based study*. *English for Specific Purposes*, 28, 183-198

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, 23(1), 35–41.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In **N. Schmitt, & M. McCarthy** (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Nattinger, J. R., & Decarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Pacquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London and New York: Continuum.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: Benjamins.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and *nativelike fluency*. In **J. C. Richards & R. W. Schmidt** (Eds.), *Language and communication* (pp. 191–226). London: Longman.

Praninskas, J. (1972). *American university word list*. London: Longman.

Shin, D & Nation, P. (2007). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*. 62/ 339-348.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.

Sinclair, J. (1966). *Beginning the study of lexis*. In C. E. Bazell, J. C. Catford, & M. A. K.

Halliday, et al. (Eds.), *In memory of J. R. Firth* (pp. 410–430). London: Longman.

- Scott, M.** (1999). *WordSmith tools users help file*. Oxford: Oxford University Press.
- Sinclair, J.** (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J.** (1998). *The lexical item*. In **E. Weigand** (Ed.), *Contrastive lexical semantics* (pp. 1–24). Amsterdam: Benjamins.
- Sinclair, J., Jones, S., & Daley, R.** (2004). *English collocation studies: The OSTI report*. London: Continuum.
- Sketch Engine.** (2018). *POS tags*. [online] Available at: https://www.sketchengine.eu/pos-tags/?utm_source=Sketch+Engine+News&utm_campaign=c378e43454-RSS_EMAIL_CAMPAIGN&utm_medium=email&utm_term=0_a385d1d459-c378e43454-155548629 [accessed on: 27/8/2018].
- Storch, N. and Tapper, J.** (2009). *The impact of an EAP course on postgraduate writing*. *Journal of English for Academic Purposes*. 8, 207-233.
- Stuart, K., & Trelis, A. B.** (2006). *Collocation and knowledge production in an academic discourse community*. [online] Available at: https://www.researchgate.net/publication/266863517_Collocation_and_knowledge_production_in_an_academic_discourse_community [accessed on: 24/8/2018]
- Stubbs, M.** (1995). Collocations and semantic profiles: On the cause of the trouble with *quantitative study*. *Functions of Language*, 2(1), 23–55.
- Stubbs, M.** (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.
- Stubbs, M.** (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
- Swales, J.** (2004). *Research genres: Explorations and applications*. Cambridge University Press.
- Teubert, W.** (2005). “*My version of corpus linguistics*”. *International Journal of Corpus Linguistics*, 10 (1), 1–13.
- Thomas, J.** (2017). *Discovering English with Sketch Engine* (2nd Edition). Versatile.

Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, 6, 18-35

University of Warwick. (2018). [online] available at: <https://warwick.ac.uk/fac/arts/> [accessed on 24/8/2018]

Wang, J., Liang and Ge, G. C. (2008). *Establishment of a Medical Academic Word List.* *English for Specific Purposes*, 27, 442-458.

West, M. (1953). *A general service list of English words.* London: Longman.

Vongpumivitch, V., Huang, J.H. & and Chang, Y.C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28, 33-41.

Xu, R., Lu, Q. & Li, Y. (2003). An automatic Chinese collocation extraction algorithm based on lexical statistics. 321 - 326. 10.1109/NLPKE.2003.1275923.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, 3, 215–299.

Appendices

Appendix 1. Overview of the study and the subset of 112 collocations

Arts and Humanities Collocations for Expert Review

As you know, I am developing a collocation list for EAP teaching-learning in Arts and Humanities. The purpose of my research is to address the need for discipline specific lists of collocations in EAP (of which there are currently none), as opposed to generic lists comparable with the AWL.

The 112 collocations you are about to review are a representative sample from a much longer list that has been derived from a corpus of PhD theses comprising (approx.) 5.5 million words. The scores from your expert review will be used as the basis for a correlation analysis in order to ascertain which qualitative or quantitative criteria are the best indicators of teaching worth, which will then allow for the weeding, collating and prioritising of items on the final list.

The collocations have been presented in context in a common inflectional and positional form. For example, the collocation *construct (v) - identity (n)* is presented in context as *construct a common identity* because in the Arts and Humanities corpus *construct* commonly occurs as the lemma with *identity* in span position + 2. However, please consider that the collocation you are reviewing also occurs in other inflectional and positional variations which would be listed under the same collocation (e.g. *constructs identities; constructing an identity; constructing identities; constructed, albeit poorly, Chinese identity – inter alia*).

Please review each collocation in consideration of these two questions:

Is it appropriate to consider the entry as an academic collocation for ESAP teaching-learning purposes?

Do you think the collocation is worth teaching explicitly as part of an Arts and Humanities ESAP course?

Once you have considered these two questions please give each collocation one score between 1 and 4 based on the following scale:

1 – definitely exclude

2 – perhaps exclude

3 – perhaps include

4 – definitely include

Thank you again for your participation,

James O'Flynn

112 collocations in context (<i>italicised and bolded</i>)	Your score (1 – 4)
Kasman notes a less overt version of this process in his observation that the piecing together of the disparate spaces of the submarine through technology in <i>Crimson Tide</i> reflects Scott's own artificial piecing together of time and space through editing.	
He was able to begin in moments of twilight, before opening them out to articulate a particular way of seeing the world.	
In this sense, then, the body of children's literature becomes a useful source material for the identification of ideology and paradigmatic shifts in social thinking	
Courtiers, merchants, humanist scholars, monks and nuns, and even Carlo Ginzburg's famous heretical miller, Menocchio, all read the text .	
However, the strategies employed by Wall in his engagement with the body are similar to my own.	
However, that reading suggests that counter readings will always be overshadowed by the dominant narrative , since frontier is a product of dominant culture and principle means of securing conquest, albeit unsuccessfully.	
The student's existing social role is challenged as he finds out how it felt for someone else.	
When the natural environment and traditional culture are destroyed, ecological ethics and local emotions become progressive thinking.	
The political messages that are inherent to the typically overlooked B-movies analysed here emerge through the application of Jameson's dialectic of artistic forms.	
When the Nurse came to the young men and asked where Romeo was, Mercutio joked that 'Tybalt killed him.	
Japan's modern period might symbolically start with the nation's declaration of modernisation, but eventually all three aspects coexisted in any one space and time.	
Mitchell and her team combined literal and abstract approaches to sound, exploring methods designed to work on the intellectual and emotional responses of the audience.	
In his book of moral rhymes, whose definitive version was printed in 1583, the poet clearly shows that he considers Horace as one of his main points of reference	
Like Genna and De Cataldo, as we shall see, they search for ways to process the past in order to move forward in the future.	
As I will illustrate in greater detail in the second chapter, is informed heavily by the genre conventions of both the western and the gothic fiction.	
The final chapter of the film begins with the modern day Orlando delivering her manuscript to a publisher.	
Every theatre performance has unique moments of improvisation on stage and unpredictable reactions from the audience.	

Indeed, as the political climate became increasingly charged in the 1960s, TeenMovies incorporated youth protest movements into their narratives.	
Even on this occasion his eyes play a crucial role insofar as they speak a thousand words.	
This however only serves to perpetuate the identification of Pulp Fiction with a depth model of hermeneutics	
Language use by individual characters is indicated in both stage directions and dialogue.	
At the same time, these new rights also force us to formulate a new model of what citizenship means.	
Moreover, the film goes to great lengths to show some of the British characters as incompetent and averse to American opinions regarding the African colonies.	
The Players in the Courtroom Examining patterns of spatial representation in the golden age courtrooms has revealed what I see as a central, structuring relationship between the court of/as law and the individuals who occupy the space .	
In addition, the aim was to strengthen the identity of the Tsou people, recover their traditional values and culture, and re-establish the social order of Tsou society.	
All these elements help also to interpret a further development in Pasinelli's career, the passage from 'traditional' production to the increasingly 'modern' one of letteratura amena.	
The male hero is almost psychologically abusive to the frail heroine.	
The production of Macbeth the Traitor clearly demonstrates that serious tragic plays may also attract audiences to the commercial theatre.	
The primary sources we used are a few letters between Madame de Flahaut, Windham and Morris kept in the British Library and the National Archives in Paris.	
Here, however, I aim to focus specifically on debates around the relationship between art and society, considering how making art might function as a practice within a society.	
Gilbert was renowned for the fastidiousness with which he staged his productions , and insisted upon their reproduction and revival with exactitude.	
The men do not look at each other's faces, nor is there much focus on their facial expressions .	
This chapter will concentrate on analysis of Potter in the broader context of an international art cinema.	
In the vast majority of cases analysed, the imitation of Horace is associated with that of other classical authors.	
He displays a strong sense of urgency to delve into his subject matter, going undercover at great risk to get under the skin of the Camorra.	
Several of these films actually chose to articulate this support via a new kind of positive representation of Britain's biggest PR liability: its imperialism.	
While on the one hand Oh seemed to punish the older generation who ruined the country, on the other he appears to bless the love of the guiltless younger generation.	

In my analyses below the changing shape of the family in early modern England is matched by modern Irish literature .	
In contrast to earlier scenes at the family home and at the police station, she stands silently to one side, indicating that she has relinquished her earlier dominance.	
Beginning with the primacy of thought and ideas, Blanqui attributes historical change to philosophical change.	
This sexual difference multiplies conceptions of temporality and spatiality.	
It built up reserves of cultural capital by challenging what counted as legitimate theatre, who were the legitimate actors, what counted as a legitimate theatre venue, and perhaps most importantly, who made up a legitimate – and reachable – theatre audience	
Hong Kong also did not play a significant role in the history of British colonialism until the last twenty years as a Crown colony.	
The rigid attitudes of critics such as Adorno, Brecht and Lukács, and the ideological reading of texts that they promote, is thus problematized.	
So far then only a small number of Korean scholars have been able to publish their research on this topic in English.	
They may not be mutually exclusive in their conceptions and meanings, but they are decidedly distinct.	
Each member of the growing band of soldiers that Bogart comes across are all stereotypes of their respective social classes and their nation.	
Next is a scene of a large group of Islamic worshippers kneeling for morning prayers; this is followed by several scenes of crowded streets, bazaars and Indian people going about their daily routine.	
The historical deposits in question constitute much of Beckett's imaginative raw material during this period.	
However, even in the early days , photographic techniques were employed to manipulate, more or less successfully, humans' perceptions of real events and environments.	
For Artaud, the mysteries were an influence on the more disturbing elements of Greek tragedy that should serve as a model for modern theatre .	
On the contrary, modern men know the truth – for example, they know that natural entities have no mind – and therefore their belief is an act of choice.	
If Suzuki's films dwell on the human experience of time , they are also interested in ethical problems.	
This 'natural animation' that is born in a regular active participation in religious rituals is something that Rilke considered Western Christianity to have lost.	
However, compared with western theories on 'multicultural citizenship', the new thinking on postmodern culture and global citizenship is still ignored by 'multicultural Taiwan'.	
The distinction between form and content , verba and res, is fundamental and absolute.	

The liminal spaces that existed between these binaries offered a degree of resistance against contemporary social and aesthetic ideas that surrounded the relationship between gender and performance.	
This chapter aims to show the beginnings of how a nomadic way of being is inscribed in the writing of Monénembo and how this prepares us to focus on Monénembo's postcolonial project.	
He begins by describing the personal experience , a walk with Sir George Beaumont around his estate at Coleorton Hall in December 1820, and their discussions regarding plans to build a new church on the site.	
It was based on a novel of the same name , written in 1933.	
Although there are no explicit references to this potential, Barbara is convinced of this fact, much to J's incomprehension.	
In this way the work is an insurgency in that it challenges the notion of normalised behaviour.	
In this very dense passage Aristotle lays down the basic principles of his linguistic theory: he draws an important boundary between things and thoughts on the one hand, and spoken and written words on the other.	
Given the long-standing literary tradition of maliciously or benevolently deceitful prophecies, this is another reason why interpreting Macbeth merely in the context of Jesuitical equivocation misses the more fundamental point.	
The world of Henry V, The Comedy of Errors, and Twelfth Night at the Watermill in the late nineties was one of a small group of young men, almost all in their late 20s and early 30s, who had little to do but "drink and rehearse."	
A key point at the outset of this trend is the replacement of Peter Parker by black Hispanic youth Miles Morales as the Ultimate universe's Spider-Man.	
In the following decade , these elements were blown out of proportion, especially in terms of their promotional use.	
Far from embodying the escape of the outcast, the yakuza, in their intractable power relations , are simply the mirror image of 'legitimate' society.	
Yet for a long time , research on Cesarotti's 'vichismo' was limited.	
Both critics explain the importance of cultural production to emerging modern nation-states.	
I will turn now to some further critical responses to these films, in order to introduce the key themes that repeatedly arise.	
A love-making scene for the main characters , with a similar metaphorical narration, is also found in King Vajiravudh's Phraya Ratchawangsan.	
During this process, on the one hand, residents can construct a common identity for their community.	
Pinky positions itself as a film about personal identity .	
Spin-offs include a graphic novel, a theatrical performance and a musical score.	
The standard German TT is very close to the original, retaining the Scottish culture, but is not able to convey the political situation in Scotland to the German audience.	

Mort's mental state is also compared with a sense of detachment and dislocation: The worst of it wasn't physical.	
The kid is a different kind of witness, a narrator of that fiction that is ethics, trying to apply it in and to his world.	
In the early stages of her career, she is a 'son-daughter' whose 'sexuality, a once frailly virginal and robustly assertive, is channelled towards the father'.	
The painting engages with the notion of intricate frameworks that construct meaning and can be understood as a visual representation of Elliott's facing mirrors.	
Thus the continuity/discontinuity debate tends to capture both sides of a reality that constituted a larger, complex social system called the postwar.	
However, the process of erosion in its different forms continually acts upon this geological process of creating and piling up.	
In chapter 6, the final content chapter, many of the issues already discussed in this chapter will be revisited via a comprehensive integration with discussions over the other chapters.	
The theatre space as a metaphor for communal everyday life is expanded to the whole narrative of the next Kihachi film.	
Summertime can be understood as a drama of 'crossing over', of coming to inhabit the space that had formally belonged only to fantasy.	
This group is not objectively constituted through its a priori inscription within the socio-economic structure but is above all created through the conscious act of political struggle .	
He spends his time inventing new ways to make fun of the monks.	
The scene of 'haunting' in Nakasago's house therefore contains a great deal of imagery that was conventional for ghost films of the 1950s.	
This remains the industry standard at the present time : all stand-bys marked in red, all 'Go's marked in green, and calls written in blue.	
This space is the product of 'lived experience, that is directly related to notions of the body and to temporality.	
A good example is a short story 'Alice in Literal-Land' by John F. Scott published in the Century Magazine in 1924.	
It is undeniable that experiencing navigable space in front of a screen operates in a different register to more traditional forms of travel.	
Obviously, the Aristotelian treatise was also used to further the discussion about new literary genres and develop new theories on them.	
He found this in a form of popular theatre usually referred to as music-hall, variety, cabaret, or café-concert.	
Africa in Sugar and Slate becomes a third party location, with a slightly different portrayal.	
In these circumstances, the ordinary distinctions between reality and fiction became blurred.	
However, they could also be found in a large number of historical novels from beyond the nebula.	

Here, though, the reporter is forced away from the scene of the crime, and gains no access to the local community who, through omertà, refuse any knowledge.	
On the one hand, the KMT government could not trust the Ben-Sheng people, so the latter could not share the same political rights as the Wei-Sheng people.	
I wish briefly to extrapolate some of the major similarities and differences identifiable across the golden age legal films.	
The sophisticated comedy may have had its roots in European operetta and Continental attitudes towards sex, but it emerged in Hollywood as a new genre .	
While I focus on a specific historical period which stretches from the end of the Second World War to the early 1970s, I do not emulate Vidotto's attempt to approach history as an evolutionary process.	
The language of an environmental critique can, in short, readily reflect a politics which threatens (directly or indirectly) the dissolution of modern life as we know it.	
Indeed, the film shows a less sympathetic view of the IRA if examined carefully.	
Along with texts written in elegiac couplets and hexameters, Mancinelli often employed schemes derived from the Horatian corpus, with a predilection for the Sapphic strophe.	
The director should have room to exercise his/her imagination to create or re-create the original text .	
Moreover, this constitutes a 'hostile theoretical recolonization' that operates to keep the Native subject at the periphery of a dominant culture .	
Stereotypical representations of Black males as rapists of white women has played a major role in the rise of racism.	
I argue below that taking Italian organized crime films as 'confrontation' of the trauma of organized crime, rather than 'compensation', facilitates this link between social history and cultural representation .	
While cultural differences in the public sphere are now supported by new forms of public policy and resources, there is also an expansion in the concept of citizenship.	
There are however aspects of Potter's persona that cannot be comfortably accommodated within this critical framework .	
These two different ways of viewing change constituted the basic narratives structure of many of Ozu's films in this period.	

Appendix 2. The Academic Collocation List for Arts and Humanities

J + N Combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCAMP	FILMTV	MODLANG	THEATRE			
active-j participation-n	5	2	3	22	32	10.82	<i>active participation of</i>
aesthetic-j experience-n	2	20	6	2	30	8.13	<i>aesthetic experience</i>
allegorical-j meaning-n	28	1	2	1	32	9.42	<i>allegorical meanings</i>
allegorical-j reading-n	41	3	1	1	46	10.17	<i>allegorical readings of</i>
basic-j principle-n	5	7	8	13	33	10.3	<i>basic principles of</i>
binary-j opposition-n	13	28	3	12	56	11.91	<i>binary oppositions</i>
broad-j context-n	4	12	4	8	28	8.56	<i>a broader context</i>
broad-j sense-n	9	8	6	13	36	9.32	<i>a broader sense</i>
capitalist-j society-n	7	6	9	6	28	8.82	<i>capitalist society</i>
central-j character-n	9	29	2	8	48	8.72	<i>central character</i>
central-j concern-n	10	9	7	3	29	9.17	<i>a central concern</i>
central-j role-n	13	17	9	6	45	9.34	<i>a central role in</i>
central-j theme-n	9	11	7	7	34	9.5	<i>central theme of</i>
certain-j aspect-n	8	10	8	2	28	8.71	<i>certain aspects of</i>
certain-j extent-n	13	4	10	5	32	9.64	<i>to a certain extent</i>
civil-j society-n	4	1	6	70	81	10.3	<i>of civil society</i>
classical-j text-n	20	2	16	7	45	8.57	<i>classical texts</i>

clear-j example-n	2	16	13	22	53	9.8	<i>a clear example of</i>
close-j analysis-n	7	51	8	4	70	10.49	<i>close analysis of</i>
close-j reading-n	11	18	14	2	45	10.04	<i>close reading of</i>
close-j relationship-n	6	13	11	25	55	9.75	<i>close relationships with</i>
collective-j identity-n	1	14	3	35	53	8.91	<i>a collective identity</i>
collective-j memory-n	3	6	30	7	46	10.1	<i>collective memory</i>
comic-j effect-n	3	2	24	1	30	9.29	<i>comic effect</i>
commercial-j success-n	6	22	11	6	45	10.83	<i>commercial success of</i>
common-j culture-n	3	7	1	19	30	7.53	<i>a common culture</i>
common-j people-n	5	5	8	11	29	8.22	<i>the common people</i>
complex-j relationship-n	6	15	7	14	42	9.24	<i>the complex relationship between</i>
contemporary-j context-n	10	6	2	10	28	7.94	<i>contemporary context</i>
contemporary-j culture-n	17	12	4	17	50	8.01	<i>contemporary culture</i>
contemporary-j society-n	22	15	12	25	74	9.33	<i>contemporary society</i>
creative-j act-n	3	1	8	18	30	8.97	<i>a creative act</i>
creative-j process-n	20	3	9	50	82	10.05	<i>the creative process</i>
critical-j analysis-n	1	13	4	12	30	8.6	<i>critical analysis of</i>
critical-j approach-n	8	15	6	3	32	8.49	<i>critical approaches</i>
critical-j attention-n	13	16	13	2	44	9.42	<i>critical attention</i>
critical-j discourse-n	6	25	3	1	35	8.6	<i>the critical discourse</i>

critical-j discussion-n	7	14	4	3	28	8.75	<i>critical discussion of</i>
critical-j distance-n	5	84	5	5	99	10.83	<i>of critical distance</i>
critical-j engagement-n	8	19	3	5	35	9.12	<i>critical engagement with</i>
critical-j framework-n	9	25	7	2	43	9.32	<i>critical framework</i>
critical-j perspective-n	2	45	13	1	61	9.64	<i>critical perspective</i>
critical-j response-n	3	17	12	3	35	9	<i>critical responses to</i>
crucial-j point-n	10	6	14	2	32	8.97	<i>a crucial point</i>
crucial-j role-n	9	5	11	4	29	9.08	<i>a crucial role in</i>
cultural-j background-n	6	1	12	18	37	7.86	<i>cultural background</i>
cultural-j capital-n	104	15	6	101	226	10.44	<i>of cultural capital</i>
cultural-j context-n	7	17	11	14	49	7.95	<i>cultural context</i>
cultural-j development-n	1	1	6	28	36	7.7	<i>cultural development</i>
cultural-j difference-n	5	8	7	201	221	10.29	<i>cultural differences</i>
cultural-j diversity-n	22	1	3	117	143	9.87	<i>of cultural diversity</i>
cultural-j element-n	2	1	23	4	30	7.27	<i>cultural elements</i>
cultural-j experience-n	7	1	1	26	35	7.29	<i>cultural experience</i>
cultural-j form-n	9	21	3	19	52	7.7	<i>cultural forms</i>
cultural-j heritage-n	9	1	7	29	46	8.22	<i>cultural heritage</i>
cultural-j history-n	10	9	15	19	53	7.96	<i>cultural history</i>
cultural-j identity-n	22	8	5	113	148	9.31	<i>cultural identity</i>

cultural-j life-n	3	2	6	38	49	7.76	<i>to participate in cultural life</i>
cultural-j memory-n	8	6	7	11	32	7.59	<i>cultural memory</i>
cultural-j practice-n	10	7	5	29	51	7.96	<i>cultural practices</i>
cultural-j product-n	2	9	9	13	33	7.7	<i>cultural products</i>
cultural-j production-n	38	13	5	31	87	8.74	<i>of cultural production</i>
cultural-j reference-n	3	1	25	2	31	7.55	<i>cultural reference</i>
cultural-j representation-n	4	11	7	9	31	7.45	<i>cultural representation</i>
cultural-j study-n	7	10	11	64	92	8.73	<i>cultural studies</i>
cultural-j tradition-n	6	5	4	43	58	8.22	<i>cultural tradition</i>
cultural-j value-n	2	19	10	20	51	8.19	<i>cultural values</i>
daily-j life-n	25	17	30	29	101	10.13	<i>of daily life</i>
dead-j body-n	6	6	11	22	45	9.8	<i>the dead body</i>
defining-v feature-n	7	16	2	19	44	9.92	<i>a defining feature of</i>
detailed-j analysis-n	4	16	9	4	33	9.67	<i>a detailed analysis of</i>
detailed-j discussion-n	13	3	10	4	30	10.02	<i>detailed discussion of the</i>
different-j approach-n	11	17	23	13	64	8.74	<i>different approaches to</i>
different-j aspect-n	4	16	8	7	35	7.98	<i>different aspects of</i>
different-j background-n	6	2	5	16	29	7.9	<i>from different backgrounds</i>
different-j character-n	7	17	9	1	34	7.45	<i>different characters</i>
different-j context-n	13	9	18	23	63	8.62	<i>different contexts</i>

different-j culture-n	17	4	11	40	72	8.23	<i>different cultures</i>
different-j form-n	16	25	27	20	88	8.71	<i>different forms of</i>
different-j genre-n	4	13	13	2	32	7.74	<i>different genres</i>
different-j group-n	4	2	12	40	58	8.49	<i>different groups</i>
different-j kind-n	26	43	9	19	97	9.65	<i>a different kind of</i>
different-j level-n	6	4	12	20	42	8.33	<i>different levels of</i>
different-j meaning-n	10	10	5	5	30	7.77	<i>different meanings</i>
different-j media-n	2	21	6	3	32	7.94	<i>in different media</i>
different-j perspective-n	8	14	13	12	47	8.43	<i>a different perspective</i>
different-j space-n	3	10	5	10	28	6.99	<i>different spaces</i>
different-j time-n	6	8	11	17	42	7.81	<i>at different times</i>
different-j version-n	10	6	10	5	31	7.86	<i>different versions of</i>
different-j way-n	28	69	42	25	164	9.64	<i>in different ways</i>
digital-j technology-n	1	29	2	24	56	10.5	<i>digital technologies</i>
domestic-j space-n	5	51	2	4	62	8.9	<i>the domestic space</i>
domestic-j sphere-n	13	7	2	9	31	10.21	<i>of the domestic sphere</i>
dominant-j culture-n	82	14	8	7	111	9.49	<i>dominant culture</i>
dominant-j ideology-n	6	28	9	3	46	10.23	<i>the dominant ideology</i>
dominant-j narrative-n	14	10	2	2	28	8.49	<i>the dominant narrative</i>
double-j meaning-n	18	1	6	3	28	9.45	<i>the double meaning</i>

early-j day-n	17	15	2	6	40	8.6	<i>the early days of</i>
early-j period-n	60	19	4	17	100	9.64	<i>in the early modern period</i>
early-j stage-n	15	7	5	28	55	8.88	<i>early stages of</i>
early-j work-n	25	30	34	6	95	9.1	<i>earlier work</i>
early-j year-n	26	39	22	15	102	9.66	<i>the early years of</i>
economic-j development-n	4	3	3	19	29	9.07	<i>economic development</i>
emotional-j response-n	12	19	8	3	42	9.97	<i>emotional response</i>
emotional-j state-n	7	26	5	5	43	9.49	<i>emotional state</i>
essential-j element-n	4	5	3	21	33	9.06	<i>an essential element of</i>
essential-j part-n	7	10	3	8	28	8.93	<i>an essential part of</i>
ethnic-j group-n	10	6	4	184	204	11.37	<i>ethnic groups</i>
everyday-j experience-n	7	13	1	19	40	8.56	<i>everyday experience</i>
everyday-j life-n	28	150	75	109	362	11.73	<i>of everyday life</i>
everyday-j practice-n	1	1	4	23	29	8.41	<i>everyday practice</i>
explicit-j reference-n	11	5	10	4	30	9.92	<i>explicit reference to</i>
extensive-j use-n	2	8	10	8	28	9.61	<i>extensive use of</i>
facial-j expression-n	2	29	1	3	35	10.77	<i>facial expressions</i>
female-j body-n	20	38	8	77	143	10.33	<i>of the female body</i>
female-j character-n	26	119	33	41	219	10.46	<i>female characters</i>
female-j figure-n	8	2	8	13	31	8	<i>female figure</i>

female-j protagonist-n	17	36	13	15	81	9.86	<i>the female protagonist</i>
female-j sexuality-n	7	14	1	17	39	9.04	<i>of female sexuality</i>
female-j voice-n	55	6	11	1	73	9.64	<i>the female voice</i>
fictional-j character-n	8	15	13	1	37	8.54	<i>a fictional character</i>
fictional-j world-n	24	12	4	4	44	9.12	<i>the fictional world of</i>
final-j chapter -n	23	14	8	13	58	9.22	<i>the final chapter</i>
final-j image-n	1	15	4	8	28	8.23	<i>final image of</i>
final-j line-n	24	6	8	3	41	9.24	<i>the final line</i>
final-j scene-n	19	43	10	9	81	9.74	<i>the final scene</i>
final-j section-n	16	25	18	11	70	10.22	<i>the section of</i>
following-j chapter-n	26	30	49	28	133	10.5	<i>in the following chapter</i>
following-j decade-n	3	14	21	1	39	8.99	<i>in the following decades</i>
following-j scene-n	4	20	2	5	31	8.37	<i>the following scene</i>
following-j section-n	15	34	24	15	88	10.02	<i>in the following section</i>
foreign-j language-n	16	1	9	6	32	8.67	<i>a foreign language</i>
further-j development-n	4	3	3	20	30	8.98	<i>further developments of the</i>
further-j discussion-n	30	10	4	5	49	9.92	<i>for further discussion of</i>
further-j evidence-n	13	2	31	4	50	10.08	<i>further evidence of</i>
further-j example-n	18	16	4	5	43	9.26	<i>further examples</i>
golden-j age-n	5	22	9	6	42	10.89	<i>the golden age</i>

good-j example-n	18	25	20	23	86	10.08	<i>a good example of</i>
good-j way-n	5	4	10	12	31	7.97	<i>the best way to</i>
grand-j narrative-n	2	4	18	6	30	8.92	<i>grand narratives</i>
great-j deal-n	15	23	26	12	76	10.32	<i>a great deal of</i>
great-j degree-n	12	6	9	3	30	9.02	<i>a greater degree of</i>
great-j detail-n	12	18	11	6	47	9.4	<i>in greater detail</i>
great-j emphasis-n	7	9	12	3	31	8.99	<i>a greater emphasis on</i>
great-j importance-n	6	8	7	10	31	9	<i>of great importance</i>
great-j length-n	5	8	13	2	28	9.05	<i>goes to great lengths to</i>
great-j success-n	4	9	10	10	33	9.04	<i>great success</i>
happy-j ending-n	27	9	5	6	47	11.99	<i>a happy ending</i>
hard-j work-n	10	13	18	4	45	8.95	<i>of hard work</i>
high-j art-n	10	6	1	15	32	9.09	<i>high art and</i>
high-j culture-n	15	4	5	24	48	8.21	<i>of high culture</i>
high-j degree-n	4	8	12	8	32	9.82	<i>a high degree of</i>
high-j level-n	3	9	6	18	36	9.43	<i>a high level of</i>
historical-j account-n	12	7	21	3	43	8.72	<i>historical account of</i>
historical-j change-n	2	2	23	2	29	7.98	<i>historical change</i>
historical-j context-n	32	31	25	23	111	9.72	<i>historical context</i>
historical-j event-n	12	38	48	15	113	9.89	<i>historical events</i>

historical-j fact-n	9	17	5	10	41	8.88	<i>historical fact</i>
historical-j moment-n	12	10	22	7	51	8.85	<i>historical moment</i>
historical-j narrative-n	4	31	10	20	65	8.9	<i>the historical narrative</i>
historical-j period-n	22	15	19	11	67	9.16	<i>historical period</i>
human-j activity-n	5	1	10	21	37	8.98	<i>human activity</i>
human-j body-n	20	22	9	21	72	9.58	<i>of the human body</i>
human-j experience-n	7	8	12	14	41	8.29	<i>of human experience</i>
human-j history-n	4	1	20	3	28	7.87	<i>of human history</i>
human-j life-n	23	2	21	10	56	8.74	<i>of human life</i>
human-j mind-n	14	3	14	6	37	9.06	<i>of the human mind</i>
imperial-j power-n	21	9	1	4	35	9.49	<i>imperial power</i>
important-j aspect-n	4	12	7	7	30	8.72	<i>an important aspect of</i>
important-j part-n	10	16	14	13	53	9.29	<i>an important part of</i>
important-j point-n	5	19	14	3	41	8.8	<i>an important point</i>
important-j question-n	8	9	5	9	31	8.85	<i>important questions</i>
individual-j character-n	6	21	6	8	41	8.41	<i>individual characters</i>
integral-j part-n	16	10	15	22	63	10.34	<i>an integral part of the</i>
key-j aspect-n	6	8	9	11	34	8.96	<i>a key aspect of</i>
key-j element-n	11	20	16	24	71	9.69	<i>a key element</i>
key-j factor-n	1	11	7	9	28	9.17	<i>a key factor in</i>

key-j feature-n	8	17	11	9	45	9.48	<i>a key feature of</i>
key-j figure-n	7	4	16	7	34	8.58	<i>a key figure</i>
key-j issue-n	7	7	6	15	35	8.94	<i>the key issues</i>
key-j moment-n	1	13	10	4	28	8.68	<i>a key moment</i>
key-j point-n	7	10	13	6	36	8.66	<i>key point</i>
key-j role-n	3	10	16	7	36	8.84	<i>a key role in</i>
large-j group-n	4	6	3	15	28	8.35	<i>a large group</i>
large-j number-n	16	19	22	20	77	10.69	<i>a large number of</i>
large-j part-n	20	9	27	4	60	9.62	<i>a large part of</i>
last-j chapter-n	23	1	8	9	41	8.95	<i>in the last chapter</i>
last-j decade-n	3	11	20	9	43	10.35	<i>the last decades of the</i>
last-j scene-n	8	9	3	10	30	8.54	<i>the last scene</i>
last-j year-n	7	5	6	10	28	8.8	<i>last years of</i>
late-j period-n	19	6	3	8	36	8.62	<i>the late Meiji period</i>
late-j work-n	14	25	19	4	62	8.8	<i>later work</i>
legal-j system-n	8	43	5	3	59	9.9	<i>the legal system</i>
liminal-j space-n	12	4	1	26	43	8.52	<i>a liminal space</i>
literary-j form-n	22	2	12	2	38	7.82	<i>literary form</i>
literary-j genre-n	12	2	32	1	47	8.79	<i>literary genres</i>
literary-j study-n	25	9	5	2	41	8.24	<i>literary studies</i>

literary-j text-n	43	2	18	5	68	8.76	<i>literary texts</i>
literary-j tradition-n	23	2	22	7	54	8.9	<i>literary tradition</i>
literary-j work-n	56	4	50	2	112	9.56	<i>literary works</i>
lived-j experience-n	17	22	9	144	192	11.69	<i>lived experience of</i>
local-j community-n	6	1	8	15	30	9.11	<i>the local community</i>
long-j history-n	18	14	4	12	48	8.96	<i>a long history</i>
long-j period-n	8	12	13	17	50	9.56	<i>a long period of</i>
long-j time-n	13	10	12	26	61	9.15	<i>for a long time</i>
long-j tradition-n	5	9	12	4	30	8.59	<i>a long tradition of</i>
main-j character-n	11	31	41	34	117	9.87	<i>the main characters</i>
main-j point-n	9	6	21	4	40	8.79	<i>main point of</i>
main-j reason-n	6	5	10	18	39	9.45	<i>the main reason for</i>
male-j character-n	27	34	14	11	86	9.44	<i>male characters</i>
male-j gaze-n	3	35	2	9	49	10.02	<i>the male gaze</i>
male-j hero-n	3	18	18	1	40	9.67	<i>the male hero</i>
male-j protagonist-n	7	31	10	7	55	9.95	<i>the male protagonist</i>
mass-j culture-n	8	14	6	118	146	9.96	<i>of mass culture</i>
mass-j media-n	3	4	8	49	64	10.78	<i>the mass media</i>
mental-j state-n	7	74	2	1	84	10.49	<i>mental states</i>
modern-j life-n	8	22	14	7	51	8.45	<i>of modern life</i>

modern-j literature-n	43	3	1	3	50	8.77	<i>modern literature</i>
modern-j man-n	3	1	22	2	28	7.64	<i>modern man</i>
modern-j period-n	54	1	2	3	60	9.18	<i>in the early modern period</i>
modern-j society-n	7	9	18	18	52	8.76	<i>modern society</i>
modern-j theatre-n	4	1	1	43	49	8.22	<i>modern theatre</i>
modern-j world-n	17	10	20	5	52	8.63	<i>in the modern world</i>
moving-j image-n	1	99	3	2	105	10.92	<i>moving images</i>
narrative-j form-n	3	19	12	3	37	8.02	<i>narrative form</i>
narrative-j structure-n	10	44	16	10	80	9.94	<i>narrative structure</i>
national-j boundary-n	4	1	1	40	46	9.65	<i>national boundaries</i>
national-j culture-n	5	1	1	51	58	8.31	<i>national culture</i>
national-j identity-n	50	27	30	220	327	11.21	<i>national identity</i>
natural-j world-n	18	5	12	8	43	8.98	<i>the natural world</i>
new-j culture-n	3	3	8	14	28	6.6	<i>a new culture</i>
new-j direction-n	9	13	4	7	33	7.63	<i>a new direction</i>
new-j form-n	30	30	44	71	175	9.39	<i>new forms of</i>
new-j generation-n	11	18	15	17	61	8.51	<i>a new generation of</i>
new-j genre-n	1	7	15	6	29	7.19	<i>the new genre</i>
new-j idea-n	5	4	9	20	38	7.67	<i>new ideas</i>
new-j identity-n	24	3	47	15	89	8.51	<i>new national identity</i>

new-j kind-n	14	20	3	8	45	8.05	<i>a new kind of</i>
new-j language-n	6	2	15	7	30	7.12	<i>a new language</i>
new-j life-n	14	8	9	11	42	7.48	<i>a new life</i>
new-j meaning-n	12	14	11	21	58	8.28	<i>new meaning</i>
new-j mode-n	6	9	6	16	37	7.7	<i>new modes of</i>
new-j model-n	6	4	7	17	34	7.49	<i>a new model of</i>
new-j nation-n	7	27	1	3	38	7.76	<i>of the new nation</i>
new-j order-n	4	9	9	11	33	7.55	<i>new order</i>
new-j perspective-n	9	9	12	13	43	7.86	<i>new perspectives</i>
new-j possibility-n	7	13	11	21	52	8.29	<i>new possibilities</i>
new-j set-n	3	7	5	18	33	7.61	<i>a new set of</i>
new-j space-n	15	6	4	29	54	7.65	<i>new space</i>
new-j structure-n	5	20	1	2	28	7.14	<i>a new structure of</i>
new-j technology-n	3	26	10	54	93	9.07	<i>new technologies</i>
new-j trend-n	3	2	8	18	31	7.56	<i>new trend</i>
new-j type-n	8	4	5	13	30	7.45	<i>a new type of</i>
new-j understanding-n	1	5	8	16	30	7.42	<i>a new understanding of</i>
new-j way-n	22	38	21	64	145	9.25	<i>new ways of</i>
new-j work-n	7	10	9	21	47	7.61	<i>new work</i>
next-j chapter-n	47	17	38	25	127	10.68	<i>in the next chapter</i>

old-j generation-n	2	15	6	14	37	9.66	<i>the older generation</i>
old-j man-n	26	19	13	14	72	9.38	<i>old man</i>
old-j woman-n	5	12	5	7	29	8.29	<i>old woman</i>
only-j way-n	34	13	13	9	69	9.2	<i>the only way to</i>
open-j space-n	9	2	1	18	30	7.94	<i>open space</i>
opening-j line-n	11	5	10	2	28	9.36	<i>the opening line</i>
opening-j scene-n	15	48	6	10	79	10.14	<i>the opening scene</i>
ordinary-j people-n	5	10	9	16	40	9.02	<i>ordinary people</i>
original-j meaning-n	9	4	12	3	28	8.74	<i>the original meaning of</i>
original-j text-n	8	4	25	31	68	9.06	<i>the original text</i>
outside-j world-n	8	19	13	7	47	9.37	<i>the outside world</i>
particular-j attention-n	6	16	9	6	37	8.9	<i>particular attention to</i>
particular-j interest-n	5	7	11	21	44	8.99	<i>is of particular interest</i>
particular-j kind-n	5	21	3	5	34	8.86	<i>a particular kind of</i>
particular-j moment-n	3	14	6	5	28	8.27	<i>particular moment</i>
particular-j way-n	6	21	4	2	33	7.85	<i>a particular way</i>
past-j event-n	2	19	9	1	31	9.13	<i>of past events</i>
personal-j experience-n	21	7	24	13	65	9.03	<i>personal experience</i>
personal-j identity-n	3	1	20	4	28	7.72	<i>personal identity</i>
personal-j memory-n	9	3	21	8	41	9.3	<i>personal memories</i>

photographic-j image-n	1	36	2	1	40	9.44	<i>the photographic image</i>
physical-j appearance-n	7	3	8	11	29	9.39	<i>physical appearance</i>
physical-j space-n	3	9	5	22	39	8.05	<i>physical space of</i>
political-j action-n	16	5	46	4	71	9.1	<i>political action</i>
political-j change-n	9	4	8	18	39	8.15	<i>political change</i>
political-j context-n	13	4	14	6	37	7.9	<i>the political context</i>
political-j discourse-n	18	4	8	8	38	8.03	<i>political discourse</i>
political-j engagement-n	5	2	32	2	41	8.44	<i>political engagement</i>
political-j event-n	6	1	4	24	35	7.94	<i>the political events</i>
political-j issue-n	8	9	16	6	39	8.16	<i>and political issues</i>
political-j message-n	6	2	38	1	47	8.74	<i>political message</i>
political-j power-n	13	4	21	11	49	8.45	<i>political power</i>
political-j reality-n	6	1	22	7	36	8.09	<i>political reality</i>
political-j right-n	1	2	4	23	30	7.8	<i>political rights</i>
political-j situation-n	10	3	14	7	34	8.17	<i>the political situation</i>
political-j struggle-n	3	3	25	4	35	8.3	<i>of political struggle</i>
political-j system-n	1	2	7	20	30	7.65	<i>the political system</i>
popular-j audience-n	6	1	4	49	60	9.27	<i>popular audiences</i>
popular-j culture-n	77	48	13	107	245	10.41	<i>popular culture</i>
popular-j theatre-n	8	1	2	32	43	8.2	<i>popular theatre</i>

postmodern-n culture-n	3	18	1	7	29	7.56	<i>postmodern culture</i>
present-j time-n	4	3	7	18	32	8.28	<i>the present time</i>
previous-j chapter-n	60	144	99	40	343	11.8	<i>in the previous chapter</i>
previous-j decade-n	4	18	7	5	34	9.52	<i>the previous decade</i>
previous-j section-n	8	26	9	7	50	9.77	<i>in the previous section</i>
previous-j work-n	6	15	11	3	35	8.15	<i>previous work</i>
primary-j source-n	3	6	1	19	29	9.56	<i>primary sources</i>
prime-j example-n	7	10	10	1	28	9.38	<i>a prime example of</i>
private-j life-n	9	16	4	7	36	8.52	<i>private life</i>
private-j space-n	7	17	4	7	35	8.06	<i>private space</i>
private-j sphere-n	24	2	1	7	34	10.29	<i>public and private spheres</i>
privileged-j position-n	8	12	8	5	33	9.67	<i>a privileged position</i>
public-j discourse-n	9	69	8	11	97	10.08	<i>the public discourse</i>
public-j performance-n	3	3	3	25	34	8.47	<i>public performance</i>
public-j space-n	9	27	11	44	91	9.16	<i>public space</i>
public-j sphere-n	21	9	10	48	88	10.64	<i>in the public sphere</i>
raw-j material-n	16	4	7	9	36	10.23	<i>raw material</i>
recent-j work-n	9	15	5	2	31	8.1	<i>recent work on</i>
recent-j year-n	33	18	23	21	95	10.49	<i>in recent years</i>
religious-j belief-n	12	4	9	4	29	9.74	<i>religious beliefs</i>

same-j name-n	15	23	5	4	47	8.57	<i>of the same name</i>
same-j period-n	9	4	18	3	34	7.83	<i>in the same period</i>
same-j year-n	27	15	46	19	107	9.48	<i>in the same year</i>
secondary-j character-n	1	16	11	3	31	8.37	<i>secondary characters</i>
selected-j text-n	61	1	3	1	66	9.4	<i>the selected texts</i>
sexual-j desire-n	5	16	3	51	75	10.67	<i>sexual desire</i>
sexual-j difference-n	3	12	2	39	56	9.89	<i>of sexual difference</i>
short-j story-n	16	4	13	1	34	9.95	<i>short stories</i>
significant-j role-n	10	7	4	12	33	8.93	<i>a significant role in</i>
similar-j vein-n	15	10	5	16	46	9.99	<i>in a similar vein</i>
similar-j way-n	23	24	17	24	88	9.44	<i>in a similar way to</i>
small-j group-n	8	12	11	23	54	9.38	<i>a small group of</i>
small-j number-n	5	12	14	15	46	10.1	<i>a small number of</i>
small-j town-n	6	20	13	5	44	10.29	<i>the small town</i>
social-j background-n	3	2	31	6	42	8.29	<i>and social background</i>
social-j change-n	22	16	32	18	88	9.15	<i>social change</i>
social-j class-n	27	15	31	29	102	9.42	<i>social class</i>
social-j condition-n	11	6	10	13	40	8.17	<i>social conditions</i>
social-j context-n	12	19	11	38	80	8.85	<i>social context</i>
social-j group-n	6	15	24	31	76	8.77	<i>social groups</i>

social-j issue-n	15	21	10	10	56	8.51	<i>social issues</i>
social-j justice-n	6	4	15	13	38	8.25	<i>and social justice</i>
social-j life-n	7	18	17	11	53	8.05	<i>of social life</i>
social-j order-n	28	21	33	26	108	9.59	<i>the social order</i>
social-j practice-n	2	4	4	22	32	7.48	<i>social practice</i>
social-j problem-n	19	49	4	12	84	9.2	<i>social problems</i>
social-j reality-n	19	15	21	22	77	9.01	<i>social reality</i>
social-j relation-n	27	9	24	14	74	8.96	<i>of social relations</i>
social-j role-n	13	5	10	10	38	7.88	<i>of social role</i>
social-j space-n	3	11	2	58	74	8.31	<i>social space</i>
social-j status-n	18	8	22	23	71	8.96	<i>social status</i>
social-j structure-n	5	21	14	16	56	8.43	<i>social structures</i>
social-j system-n	6	12	4	12	34	7.67	<i>social system</i>
social-j value-n	1	8	8	13	30	7.65	<i>social values</i>
social-j world-n	2	27	9	12	50	8.07	<i>the social world</i>
special-j issue-n	12	14	12	3	41	9.62	<i>special issue of</i>
specific-j context-n	4	14	13	19	50	8.98	<i>specific context</i>
stark-j contrast-n	11	22	14	12	59	12.35	<i>in stark contrast to the</i>
strong-j sense-n	14	12	17	11	54	9.79	<i>strong sense of</i>
subject-j position-n	10	14	1	13	38	9.64	<i>subject position</i>

subjective-j experience-n	18	26	11	5	60	9.39	<i>subjective experience</i>
textual-j analysis-n	11	129	24	2	166	11.64	<i>textual analysis</i>
theatrical-j performance-n	6	2	2	59	69	9.75	<i>theatrical performance</i>
thematic-j concern-n	7	23	5	3	38	10.21	<i>thematic concerns</i>
traditional-j culture-n	4	5	4	37	50	8.14	<i>traditional culture</i>
traditional-j form-n	4	7	13	16	40	8.08	<i>traditional forms of</i>
traditional-j value-n	2	8	18	13	41	9.12	<i>traditional values</i>
traditional-j view-n	12	1	1	16	30	8.78	<i>the traditional view</i>
urban-j space-n	3	15	7	28	53	8.64	<i>urban space</i>
various-j form-n	15	25	33	17	90	9.27	<i>various forms of</i>
various-j way-n	16	19	15	15	65	9.03	<i>in various ways</i>
vast-j majority-n	6	15	15	2	38	11.61	<i>the vast majority of</i>
visual-j image-n	3	15	1	10	29	8.4	<i>the visual image</i>
visual-j representation-n	5	28	4	3	40	9.28	<i>visual representation of</i>
white-j male-n	11	23	1	1	36	9.87	<i>white male</i>
white-j man-n	34	57	1	5	97	9.82	<i>white man</i>
white-j woman-n	16	32	1	2	51	9.12	<i>white woman</i>
wide-j context-n	9	5	9	11	34	8.7	<i>the wider context</i>
wide-j variety-n	12	3	8	8	31	9.7	<i>a wide variety of</i>
written-j word-n	11	65	10	4	90	10.79	<i>the written word</i>

young-j boy-n	2	14	10	4	30	9.03	<i>young boys</i>
young-j generation-n	12	16	6	27	61	9.91	<i>the younger generation</i>
young-j girl-n	15	7	17	9	48	9.55	<i>a young girl</i>
young-j man-n	42	127	82	19	270	11.07	<i>young men</i>
young-j people-n	16	18	21	157	212	10.79	<i>the young people</i>
young-j woman-n	67	70	22	8	167	10.57	<i>young woman</i>

N + N Combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCAMP	FILMTV	MODLANG	THEATRE			
actor-n (and/or) audience-n	8	2	4	15	29	9.67	<i>actors and the audience</i>
actor-n (and/or) director-n	15	9	19	9	52	10.71	<i>actor and director</i>
analysis-n (of) text-n	6	14	12	2	34	9.06	<i>analysis of the text</i>
art-n (and/or) culture-n	9	11	1	20	41	9.16	<i>art and culture</i>
audience-n member-n	15	24	7	320	366	12.54	<i>audience members</i>
body-n (and/or) space-n	1	20	3	26	50	9.78	<i>body and space</i>
body-n (of) work-n	13	15	8	32	68	10.5	<i>body of work</i>
centre-n stage-n	16	5	6	7	34	10	<i>centre stage</i>
chapter-n (of) thesis-n	4	8	25	5	42	11.5	<i>chapter of this thesis</i>

construction-n (of) identity-n	7	5	11	35	58	10.32	<i>construction of an identity</i>
culture-n (and/or) identity-n	10	7	1	49	67	9.75	<i>culture and identity</i>
culture-n (and/or) tradition-n	8	2	4	17	31	8.94	<i>culture and traditions</i>
experience-n (of) time-n	14	2	7	6	29	8.85	<i>experience of time</i>
family-n home-n	9	21	4	7	41	10.02	<i>the family home</i>
family-n life-n	3	21	7	3	34	8.35	<i>family life</i>
family-n member-n	14	12	5	23	54	9.91	<i>family members</i>
film-n industry-n	2	88	11	4	105	9.77	<i>the film industry</i>
film-n study-n	1	41	5	1	48	8.16	<i>film studies</i>
film-n (and/or) series-n	2	50	2	3	57	9.5	<i>film and television series</i>
form-n (and/or) content-n	20	36	23	11	90	11.34	<i>form and content</i>
form-n (of) art-n	14	4	6	9	33	8.5	<i>form of art</i>
form-n (of) expression-n	3	19	10	16	48	9.31	<i>form of expression</i>
form-n (of) theatre-n	34	1	3	4	42	8.74	<i>form of theatre</i>
gender-n (and/or) class-n	4	21	1	29	55	10.82	<i>gender and class</i>
gender-n identity-n	4	46	4	3	57	8.98	<i>gender identity</i>
gender-n role-n	4	28	26	9	67	9.98	<i>of gender roles</i>
memory-n (and/or) history-n	4	14	15	9	42	9.66	<i>memory and history</i>
hotel-n room-n	3	13	22	1	39	11.33	<i>hotel room</i>
image-n (of) woman-n	3	23	5	13	44	9.6	<i>the image of a woman</i>

interpretation-n (of) text-n	17	4	11	1	33	9.28	<i>interpretation of a text</i>
literature-n review-n	3	31	5	5	44	10.75	<i>the literature review</i>
love-n story-n	4	7	11	10	32	9.76	<i>a love story</i>
media-n text-n	1	38	1	1	41	8.57	<i>media texts</i>
mode-n (of) representation-n	11	14	5	5	35	10.13	<i>a mode of representation</i>
novel-n (and/or) film-n	5	11	13	1	30	8.4	<i>novel and film</i>
object-n (of) study-n	7	9	10	7	33	10.38	<i>an object of study</i>
part-n (of) chapter-n	32	10	21	5	68	9.64	<i>part of this chapter</i>
past-n (and/or) present-n	7	18	10	16	51	11.78	<i>the past and the present</i>
point-n (of) departure-n	5	10	16	9	40	10.25	<i>point of departure</i>
point-n (of) reference-n	10	15	35	3	63	10.73	<i>point of reference</i>
portrayal-n (of) life-n	3	2	27	1	33	9.32	<i>portrayal of life</i>
power-n relation-n	13	26	6	13	58	10.5	<i>power relations</i>
power-n structure-n	9	12	5	3	29	9.08	<i>power structures</i>
production-n (of) play-n	6	1	8	17	32	9.61	<i>production of the play</i>
reading-n (of) text-n	4	20	18	1	43	9.61	<i>reading of the text</i>
reality-n (and/or) fiction-n	9	3	17	6	35	10.37	<i>between reality and fiction</i>
reference-n point-n	5	18	45	14	82	10.47	<i>reference point for</i>
representation-n (of) character-n	4	17	10	2	33	8.99	<i>representation of the character</i>
representation-n (of) woman-n	2	24	4	13	43	9.56	<i>representation of women</i>

research-n question-n	2	6	10	22	40	9.85	<i>research questions</i>
section-n (of) chapter-n	21	34	16	4	75	11.6	<i>section of this chapter</i>
culture-n (and/or) society-n	8	6	7	15	36	9.12	<i>culture and society</i>
sound-n (and/or) image-n	2	36	4	6	48	10.58	<i>sound and image</i>
source-n material-n	9	126	10	6	151	11.76	<i>source material</i>
source-n text-n	10	24	139	4	177	10.66	<i>of the source text</i>
stage-n performance-n	1	1	1	25	28	8.18	<i>stage performance</i>
television-n drama-n	8	52	1	13	74	10.07	<i>television drama</i>
television-n series-n	7	100	4	10	121	11.02	<i>television series</i>
theatre-n performance-n	2	5	2	26	35	8.25	<i>theatre performance</i>
theatre-n space-n	4	1	1	38	44	7.95	<i>theatre space</i>
theology-n (and/or) philosophy-n	6	1	19	2	28	11.1	<i>theology and philosophy</i>
thought-n process-n	3	22	2	7	34	9.31	<i>thought processes</i>
time-n (and/or) space-n	38	83	20	159	300	11.93	<i>time and space</i>
time-n period-n	23	9	14	2	48	9.57	<i>time period</i>
use-n (of) term-n	12	23	13	22	70	9.95	<i>use of the term</i>
use-n (of) music-n	6	26	5	1	38	9.13	<i>use of music</i>
word-n (and/or) image-n	2	58	1	3	64	10.46	<i>words and images</i>

(SUBJ.) N + V Combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCOMP	FILMTV	MODLANG	THEATRE			
actor-n play-v	14	12	7	34	67	10.3	<i>actor playing</i>
audience-n see-v	12	12	2	14	40	8.98	<i>audience sees</i>
chapter-n aim-v	6	6	6	10	28	9.51	<i>this chapter aims to</i>
chapter-n demonstrate-v	7	13	6	4	30	9.08	<i>chapter demonstrates</i>
chapter-n examine-v	19	20	13	12	64	10.68	<i>chapter examines the</i>
chapter-n explore-v	9	26	8	11	54	10.32	<i>chapter explores the</i>
chapter-n focus-v	8	14	8	6	36	9.67	<i>chapter focuses on</i>
film-n show-v	1	29	4	1	35	7.94	<i>the film shows</i>

V + (OBJ.) N combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCOMP	FILMTV	MODLANG	THEATRE			
add-v emphasis-n	17	11	3	20	51	11.76	<i>emphasis added</i>
blur-v line-n	4	11	9	4	28	10.4	<i>blurring the lines</i>
challenge-v notion-n	6	8	5	14	33	9.7	<i>challenge the notion</i>
construct-v identity-n	8	7	4	37	56	9.73	<i>construct a new national identity</i>
construct-v image-n	3	16	7	3	29	8.75	<i>constructed image</i>

construct-v space-n	3	14	1	45	63	9.58	<i>constructed space</i>
create-v sense-n	16	32	13	14	75	9.29	<i>creating sense</i>
create-v space-n	24	20	9	69	122	9.89	<i>create a space</i>
demonstrate-v way-n	4	18	3	4	29	8.38	<i>demonstrate the ways</i>
draw-v parallel-n	17	11	15	3	46	10.27	<i>draws parallels</i>
employ-v strategy-n	3	22	3	17	45	10.14	<i>strategies employed</i>
employ-v term-n	14	5	4	6	29	9.32	<i>employ the term</i>
encourage-v audience-n	4	13	3	8	28	9.18	<i>audience is encouraged</i>
explore-v issue-n	9	12	4	10	35	9.23	<i>explore issues</i>
explore-v relationship-n	15	11	2	6	34	9.06	<i>explore the relationship</i>
explore-v way-n	13	32	4	4	53	9.15	<i>explore ways</i>
follow-v model-n	5	3	28	3	39	8.79	<i>following the model</i>
give-v sense-n	14	13	23	14	64	8.84	<i>give a sense</i>
give-v voice-n	12	10	22	8	52	8.9	<i>give voice to</i>
inhabit-v space-n	2	13	8	5	28	8.96	<i>inhabit space</i>
make-v argument-v	10	33	3	2	48	8.12	<i>arguments made</i>
occupy-v space-n	6	32	2	15	55	9.83	<i>space occupied by</i>
offer-v example-n	13	14	22	10	59	9.39	<i>offers an example</i>
offer-v insight-n	11	13	14	14	52	9.67	<i>offers an insight</i>
offer-v perspective-n	6	18	9	8	41	9.18	<i>offers a good perspective</i>

perform-v role-n	18	21	3	8	50	9.65	<i>performing the role of</i>
play-v character-n	5	15	1	9	30	8.57	<i>character played by</i>
portray-v character-n	8	12	4	5	29	8.99	<i>character portrayed by</i>
produce-v film-n	1	49	19	1	70	9.41	<i>films produced by</i>
produce-v text-n	14	11	23	1	49	9.15	<i>texts produced by</i>
publish-v text-n	11	2	19	3	35	9.06	<i>texts published</i>
read-v text-n	22	8	12	3	45	9.3	<i>read the text</i>
shift-v focus-n	9	7	5	9	30	10.46	<i>shift the focus</i>
shoot-v film-n	5	17	3	4	29	8.77	<i>film was shot</i>
stage-v play-n	7	1	5	30	43	10.09	<i>play was staged</i>
stage-v production-n	7	1	2	23	33	10.04	<i>production was staged</i>
support-v argument-n	7	6	3	21	37	10.19	<i>support the argument</i>
use-v image -n	8	13	6	7	34	7.82	<i>images used</i>
use-v metaphor-n	12	16	7	2	37	8.25	<i>metaphors used</i>
use-v strategy-n	2	21	9	16	48	8.52	<i>strategies used</i>
write-v text-n	5	6	30	5	46	8.96	<i>texts written</i>

A + J combinations	Raw freq. in each subcorpus:		logDice	Longest commonest match
---------------------------	-------------------------------------	--	----------------	--------------------------------

	ENGCOMP	FILMTV	MODLANG	THEATRE	Raw freq. in AHC		
completely-a different-a	6	5	18	6	35	10.32	<i>a completely different</i>
culturally-a specific -j	2	6	1	20	29	11.96	<i>culturally specific</i>
even-a great-j	10	9	7	7	33	10.73	<i>an even greater</i>
far-a great-j	6	13	10	1	30	11.5	<i>to a far greater</i>
highly-a influential-j	6	4	6	12	28	10.49	<i>highly influential</i>
particularly-a evident-j	1	9	19	2	31	10.08	<i>this is particularly evident</i>
particularly-a important-j	8	9	11	9	37	10.04	<i>is particularly important</i>
particularly-a interesting-j	10	19	35	5	69	11.19	<i>is particularly interesting</i>
particularly-a relevant-j	1	8	14	6	29	9.96	<i>particularly relevant to</i>
mutually-a exclusive-j	8	17	10	6	41	13.25	<i>are not mutually exclusive</i>
slightly-a different-j	12	21	14	2	49	10.82	<i>a slightly different</i>
well-a aware-j	5	10	6	10	31	11.78	<i>well aware of the</i>

A + V combinations	Raw freq. in each subcorpus:				Raw freq. in AHC	logDice	Longest commonest match
	ENGCOMP	FILMTV	MODLANG	THEATRE			
already-a discuss-v	16	8	8	1	33	8.82	<i>already discussed</i>

already-a know-v	22	9	4	5	40	8.97	<i>already know</i>
already-a mention-v	18	3	25	12	58	9.72	<i>already mentioned</i>
already-a see-v	20	5	27	9	61	9.14	<i>we have already seen</i>
bring-v together-a	31	16	21	15	83	10.97	<i>brings together</i>
clearly-a define-v	14	12	13	6	45	9.94	<i>clearly defined</i>
clearly-a demonstrate-v	4	8	12	9	33	9.45	<i>clearly demonstrates</i>
clearly-a indicate-v	4	2	4	23	33	9.6	<i>clearly indicates that</i>
clearly-a see-v	3	8	7	12	30	8.41	<i>most clearly seen</i>
clearly-a show-v	5	4	14	10	33	9.34	<i>clearly shows</i>
closely-a associate-v	7	16	3	8	34	10.49	<i>closely associated with the</i>
closely-a relate-v	7	20	15	6	48	10.87	<i>closely related to</i>
come-v together-a	18	11	8	19	56	10.11	<i>come together</i>
commonly-a use-v	8	5	14	9	36	9.86	<i>commonly used in</i>
directly-a address-v	6	11	9	3	29	9.89	<i>directly addressed</i>
directly-a link-v	2	11	11	6	30	9.7	<i>directly linked to</i>
directly-a relate-v	9	12	8	14	43	10.29	<i>directly related to</i>
draw-v together-a	6	13	7	7	33	9.66	<i>drawn together</i>
explicitly-a state-v	11	11	9	1	32	10.29	<i>explicitly stated</i>
focus-v specifically-a	5	8	8	8	29	10.2	<i>focus specifically on</i>
fully-a understand-v	13	10	2	9	34	10.15	<i>to fully understand</i>

go-v far-a	18	16	9	5	48	10.36	<i>go so far</i>
immediately-a follow-v	9	11	11	4	35	10.38	<i>immediately followed by</i>
become-v increasingly-a	8	10	9	11	38	10.08	<i>became increasingly</i>
inextricably-a link-v	3	11	5	11	30	10.99	<i>inextricably linked to the</i>
instead-a focus-v	11	9	7	6	33	9.96	<i>instead focus</i>
mainly-a focus-v	3	4	19	4	30	10.26	<i>mainly focused on</i>
move-v forward-a	4	9	9	6	28	10.47	<i>to move forward</i>
note-v here-a	17	8	3	10	38	9.15	<i>to note here that</i>
only-a exist-v	17	6	4	19	46	8.44	<i>only exist</i>
only-a make-v	18	13	7	7	45	8.26	<i>only made</i>
only-a see-v	14	25	5	6	50	8.18	<i>only see</i>
only-a serve-v	8	16	14	5	43	8.38	<i>only serves to</i>
originally-a publish-n	7	16	6	1	30	10.85	<i>originally published in</i>
previously-a discuss-v	14	18	2	3	37	10.12	<i>previously discussed</i>
primarily-a concern-v	8	14	2	5	29	10.62	<i>primarily concerned with the</i>
see-v here-a	10	2	9	15	36	8.38	<i>seen here</i>
soon-a become-v	14	6	11	8	39	10.16	<i>soon became</i>
still-a hold-n	7	6	5	11	29	8.82	<i>still held</i>
still-a remain-n	17	12	7	5	41	9.24	<i>still remains</i>
take-v together-a	10	6	8	7	31	9.19	<i>taken together</i>

trace-v back-a	21	4	16	7	48	10.46	<i>can be traced back to the</i>
well-a receive-v	2	7	2	18	29	9.67	<i>well received by</i>
work-v together-a	12	21	7	30	70	10.63	<i>work together</i>

Application for Ethical Approval

BA/MA/MSc Students

We are committed to ensuring that all research undertaken by our members, staff and students, meets the highest possible ethical standards. You will already have been introduced to research ethics in your research methods modules, but now that you are about to embark on a research project it is essential that you consider very carefully the ethical issues that it might raise and that you discuss these with your supervisor. Please treat this not only as a means of ensuring that your research meets appropriate ethical standards but also as a learning opportunity.

INSTRUCTIONS FOR STUDENTS:

Please complete PART 1 (sections A–F) and email the form to your project supervisor, together with any participant consent forms you plan to use

PART 1 (for completion by student)

A: YOUR DETAILS

<i>Student name:</i>	James O'Flynn
<i>University ID number:</i>	1766500
<i>Degree programme:</i>	MA ELT (with a specialism in EAPP)
<i>Provisional project title:</i>	Developing an academic collocation list for the hard sciences
<i>Supervisor name:</i>	Sue Wharton

B: TYPES OF DATA TO BE COLLECTED

Please describe the types of data you plan to collect (e.g. data from questionnaires, interviews, observations, conversations, experiments, media texts, images, websites, social media posts, etc.)

I will be collecting MA dissertations or PHD theses, exam papers and journal articles.

There will be a questionnaire type document on which participants will have to grade collocations from 1 (not useful) to 4 (very useful).

Are the data in the public domain?

YES/NO

If NO, explain what steps you will take to obtain permissions for data collection and use (from research participants, social media account holders, etc.)

The dissertations/ theses, exam papers, journals will be in the public domain.

Permission for the questionnaire type data will be gained and documented through consent forms.

C: PARTICIPANTS

Please describe the participants in the research (including ages of children or young participants where appropriate). Please specify if any participants are vulnerable (e.g. with a learning disability, in medical care, or in a dependent or unequal relationship; discuss with your supervisor if uncertain):

No children, no venerable people.

Participants will be academics.

Please explain what steps you will take to ensure that the fundamental rights and dignity of participants will be respected (e.g. confidentiality, privacy, anonymity, cultural or religious values):

All participants' academic roles will be anonymised. All data will be stored on a secure computer. No information about the participants, apart from their academic role, will be obtained – it isn't even necessary to collect their names.

Please indicate whether you have an existing relationship with research participants (e.g. teacher–student, employer–employee), and if so, what implications this may have for them:

There may be a teacher-student relationship with some of the research participants. Most of the participants will have no prior relationship with me – they will be contacted by email by me specifically for the purposes of the study. There are no implications for them, they just have to write 1-4 on a piece of paper. Their identities will be anonymised.

D: CONSENT

Will prior informed consent be obtained from participants?

YES/NO

If YES, explain how you will obtain consent, and whether consent will be written or verbal.

In NO, give reasons for this:

Yes, I will gain consent first by email and then using a consent form. There will be a signed form and email record of consent. The consent form and emails will detail how the data will be used, and state that participants can withdraw at any time.

<i>Will prior informed consent be obtained from others (e.g. parents/guardians, gatekeepers)?</i>	YES/NO
<i>If YES, explain how you will obtain consent, and whether consent will be written or verbal: In NO, give reasons for this:</i>	
No. They are academics, I don't need to inform their parents, guardians or any gatekeepers.	
<i>Will participants be informed of your status/role as a student researcher?</i>	YES/NO
<i>Will any form of deception be used?</i>	YES/NO
<i>If YES, explain why deception is necessary, and whether and how you will debrief the participants:</i>	
<i>Will participants be told that they can withdraw from the study at any time?</i>	YES/NO
<i>Will participants be informed of the use to which data will be put?</i>	YES/NO
<i>Will confidentiality of data be guaranteed?</i>	YES/NO
<i>If YES, what steps will you take to ensure data confidentiality? If NO, how will you negotiate this with participants before obtaining consent?</i>	
Each paper will be marked with a code that is specific to the participant's job role before distributing it to them, for example, if the participant is a physics lecturer, the code PL could be used. This way, only I know who completed the form. The participant does not have to mark or give any other information, other than the numbers 1 (not useful) to 4 (useful), i.e. there is no personal information.	

Please attach any consent forms you will be using when you email this application to your supervisor

E: SECURITY AND PROTECTION

Describe the nature and degree of any potential risk (physical, or psychological/emotional, such as reference to personally sensitive issues) to participants and what steps will be taken to deal with this:

There is no potential risk to any participants.

Describe the nature and degree of any potential risk (physical, psychological, emotional) to you as researcher and what steps will be taken to deal with this:

There is no potential risk.

Where and how long will data be stored and what measures will be taken to ensure security?

Data will be stored on a secure computer for as long as it is needed for research purposes. It will then be safely destroyed.

F: DECLARATION AND SIGNATURE

I confirm that I have read the University Statement of the Ethical Conduct of Research (http://www2.warwick.ac.uk/services/ris/research_integrity/code_of_practice_and_policies/statement_ethical_conduct_research) and the BAAL Recommendations for Good Practice in Applied Linguistics Student Projects (https://baalweb.files.wordpress.com/2017/08/goodpractice_stud.pdf):

<i>Signature:</i>	James O'Flynn
<i>Date:</i>	31/3/2018

PART 2 (for completion by project supervisor)

<i>Supervisor name:</i>	Sue Wharton
<i>Student name:</i>	James O'Flynn
<i>Have you discussed the ethical issues relating to this project with the student?</i>	YES/
<i>Will the project entail working with children or vulnerable adults?</i>	/NO
<i>Will DBS (CRB) checks be needed?</i>	/NO
<i>Will the project involve sensitive data that may be stressful for participants?</i>	/NO
<i>Will the project entail potential significant risks for participants and/or student?</i>	/NO
<i>Please comment on any issues raised above or concerns you may have:</i>	
<i>Signature:</i>	<i>SMWharton</i>
<i>Date:</i>	16 April 2018

PART 3 (for completion by Course Manager or nominee, or, where relevant, by CAL Student Research Ethics Committee Chair)

Action taken (X)

<i>Approved:</i>	Yes
<i>Approved with modifications or conditions noted below:</i>	
<i>Action deferred, with reasons noted below:</i>	

<i>Signature:</i>	AMPinter	<i>Date:</i>	19/04/2018
-------------------	----------	--------------	------------

